CONSISTENT SUBSPACE AGGREGATION FOR PERSONALIZED FEDERATED LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated fine-tuning with LoRA suffers from an aggregation inconsistency, as averaging low-rank factors does not recover the average of the updates they induce. Existing remedies need both factors at the server, but personalized protocols share only the right factors and keep client-specific left factors local, where factor averaging depends on each client's arbitrary factorization. We propose PERSIA, a fully low-rank aggregation rule for this setting, where each client whitens its right factor with the Gram matrix of its private left factor, and the server recovers a shared right subspace by truncated SVD. We prove that the core PERSIA rule exactly minimizes the clients' reconstruction error among all shared rank-r right factors, with an objective invariant to equivalent factorizations, and give a factor-stationarity bound whose floor does not grow with the number of local steps. On GLUE with RoBERTa-Large, right after aggregating a 1280-step local phase, PERSIA's personalized models are as accurate as the clients' local models (91.1% versus 90.8%), whereas factor averaging loses 4.1 points and two-factor FedAvg 11.3 points; after further rounds, final accuracy is on par across rules, with PERSIA uploading half as much as two-factor sharing.

1 INTRODUCTION

Recent advances in Foundation Models (FM) have demonstrated exceptional performance across diverse downstream applications, benefiting from massive-scale computational and data resources (Liu et al., 2019; Brown et al., 2020; Li et al., 2023; Liu et al., 2023; DeepSeek-AI, 2024). However, full-rank fine-tuning of FMs remains computationally expensive and data-intensive for users with limited resources. To address this, Parameter-Efficient Fine-Tuning (PEFT) approaches, notably Low-Rank Adaptation (LoRA) (Hu et al., 2022), have emerged as practical solutions to reduce computational requirements (Xu et al., 2026). While LoRA considerably alleviates fine-tuning overhead, individuals with limited data struggle to adapt these models to their specific tasks effectively (Mao et al., 2025).

Federated Learning (FL) enables clients to collaboratively fine-tune models on local datasets without sharing their data. In FL, clients periodically train and upload model parameters to a central server, which aggregates them and broadcasts the aggregated model back to the clients (McMahan et al., 2017). Because client data are heterogeneous, a single global model rarely fits every client, which motivates personalized FL (PFL), where clients share part of the model and keep the rest local (Smith et al., 2017; Tan et al., 2023). For LoRA, FedSA-LoRA (Guo et al., 2025) shares the right factor A and keeps the left factor B on each client, following the observation that A captures common knowledge while B adapts to the local data distribution (Zhu et al., 2024).

Challenges. Despite LoRA's appeal for federated fine-tuning, the aggregation is nontrivial. Naive factor-wise averaging of LoRA parameters suffers from the *low-rank aggregation problem*:

$$\frac{1}{N}\sum_{i=1}^{N}(B_iA_i) \neq \big(\frac{1}{N}\sum_{i=1}^{N}B_i\big)\big(\frac{1}{N}\sum_{i=1}^{N}A_i\big)\ . \tag{1}$$

The aggregated adapters deviate from the full-rank update, leading to inconsistencies between local and global steps. Workarounds using full-rank reconstruction require both factors at the server. In personalized FL, where A is shared globally while B remains local, the server does not have access to both factors to perform full-rank reconstruction, as shown in Figure 1, analogous to the FedSA-LoRA (Guo et al., 2025) methodology. Although no left factors are averaged in this case, the inconsistency

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

persists in a less visible form. LoRA is non-identifiable, $BA = (BT)(T^{-1}A)$ for every invertible $T \in \mathbb{R}^{r\times r}$, so the coordinates of a client's A_i are meaningful only relative to its private B_i.

As a result, federated fine-tuning with LoRA often relies on restrictive local training, including fewer local steps, smaller learning rates, and client drift regularization, slowing convergence and increasing communication to control aggregation error. These limitations are especially costly when communication is scarce, since each aggregation must preserve as much locally learned information as possible. This tension between consistency and scalability motivates the following question:

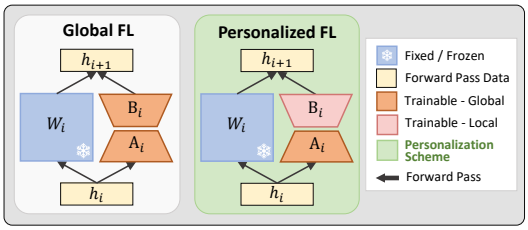


Figure 1: A single-layer diagram of LoRA in global FL and FedSA-LoRA in personalized FL. In the personalized protocol, the server never receives B_i.

Can we recover a consistent shared LoRA subspace without accessing clients' private left factors, while remaining effective under long local training intervals and sparse communication?

We introduce PERSIA (PErsonalized Right-Subspace Invariant Aggregation), a novel low-rank aggregation rule that learns a shared low-rank *right subspace* by minimizing the mean-squared residual of right projections across all client updates. The key ideas of our method are as follows:

(i) To ensure a metric-correct right subspace, we apply Gram whitening, which encodes the right-side metric of each update in a right-factor-sized message: $Z_i = (B_i^\top B_i)^{1/2} A_i$ satisfies $Z_i^\top Z_i = (B_i A_i)^\top (B_i A_i)$, so it carries the right-side geometry of the update and none of its arbitrary coordinates.

(ii) We compute the top r principal components from all clients' whitened right factors via SVD and rebase clients' left factors onto the shared subspace to ensure consistency before aggregation.

(iii) In personalized FL, whitening and rebasing happen locally, so B_i never leaves the client, and **(iv)** all aggregation operates on low-rank factors, with time and memory linear rather than quadratic in the model width.

Contributions. We make the following contributions:

- We present a principled aggregation method that mitigates the low-rank averaging problem when only the right factors are shared with the server, thereby enabling personalization alignment within the FedSA-LoRA architecture, with complexity linear in the layer width.
- We prove that PERSIA exactly minimizes the mean client reconstruction error among *all* shared rank-r right factors, with an objective invariant to equivalent LoRA factorizations, so it never reconstructs the same client updates worse than factor averaging or unwhitened alternatives.
- We establish a factor-stationarity bound for the core rule, in the factor variables that the algorithm updates, with an explicit floor that does not grow with the number of local steps.
- On five GLUE tasks with RoBERTa-Large, we measure what each aggregation preserves: right after aggregating up to 1280 local steps, PERSIA's personalized models stay within two points of the clients' local models and exceed them on average at $\tau = 1280$, whereas factor averaging and two-factor averaging lose up to 10 and 26 points. Further rounds of local training repair these losses, and final accuracy is on par, with PERSIA matching the most accurate baseline at half the uplink of two-factor sharing; Gram whitening is essential.

An overview of PERSIA is illustrated in Figure 2. Each plane corresponds to a client-specific r-dimensional subspace spanned by the columns of B_i, and the columns of A_i are coordinates within it. Gram whitening makes each update's right-side metric independent of its factorization, the SVD recovers a shared right factor under Euclidean geometry, and rebasing expresses each client's left factor in this shared basis.

2 Related Work

LoRA in Federated Learning. Following LoRA's success in centralized fine-tuning, various studies have integrated LoRA into FL to improve communication efficiency (Fang et al., 2026), privacy

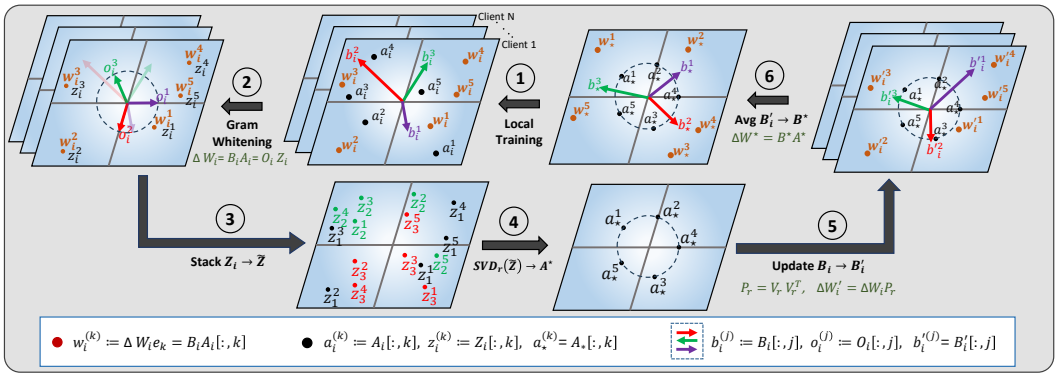


Figure 2: **PERSIA Overview.** (1) Clients train locally. (2) Clients apply Gram-whitening to remove local geometric distortions (rotation/scaling) induced by the left factor. (3) Server stacks all whitened factors on a plane. (4) Server performs a truncated SVD to find the shared right subspace. (5) Clients rebase their left factors into this shared coordinate system, which completes the personalized aggregation. (6) Only in the global-FL variant (Appendix E), the server additionally averages the rebased left factors and broadcasts them to clients.

preservation (Sun et al., 2024), and personalization (Guo et al., 2025). FedSA-LoRA builds on the LoRA asymmetry (Zhu et al., 2024): A tends to capture *common knowledge*, while B adapts to the data distribution and encodes *personalized preferences*. However, factor-wise averaging fails to capture the full-rank update that LoRA represents, which induces a functional misalignment between what clients learn and what the server broadcasts.

Exact and two-factor aggregation. Several recent methods remove the mismatch in Equation (1) when both factors reach the server: FedEx-LoRA (Singhal et al., 2025) adds the aggregation residual to the frozen weights, FLoRA (Wang et al., 2024) stacks the client adapters and merges them into the base model, LoRA-FAIR (Bian et al., 2025) refines the averaged factors at the server, and FLoRIST (Ramesh & Dass, 2026) thresholds the singular values of the stacked factors. Fed-SB (Singhal et al., 2026) and Ravan (Raje et al., 2025) train small components between frozen outer factors, and FFA-LoRA (Sun et al., 2024) freezes A. All of these require the left factors at the server or change the trainable parameterization, so none applies when a trainable B_i stays on the client; Appendix E compares with all of them in the two-factor setting except Fed-SB and Ravan, which change the trained parameterization.

Personalized FL. Personalized FL tailors models to heterogeneous clients through multi-task formulations (Smith et al., 2017), regularization toward a global model (Dinh et al., 2020; Li et al., 2021), or shared representations with client-specific heads (Arivazhagan et al., 2019; Collins et al., 2021). For foundation models, FedDPA (Yang et al., 2024) and FDLoRA (Lu et al., 2024) combine global and local adapters, PF2LoRA (Hao et al., 2025) learns two levels of LoRA with automatic rank selection, HiLoRA (Peng et al., 2026) organizes root, cluster, and client adapters, FedALT (Bian et al., 2026) mixes a local adapter with a rest-of-world adapter, and tri-matrix adaptation communicates a small core between the two factors (Li et al., 2025). In most of these designs the shared component is a complete adapter aggregated by the server, so an aggregation rule such as PERSIA is complementary. FedSA-LoRA (Guo et al., 2025) instead keeps one factor of the *same* adapter local, precisely where server-side reconstruction is impossible.

LoRA Alignment. A classic approach to improving stability is to regularize the local objective, as in FedProx (Li et al., 2020) and SCAFFOLD (Karimireddy et al., 2020). These methods are orthogonal to the LoRA averaging issue, as they reduce parameter-space drift without making LoRA factors comparable across clients. Weight alignment methods handle permutation symmetries by matching hidden units before averaging, including FedMA (Wang et al., 2020a), Git Re-Basin (Ainsworth et al., 2023), and related Optimal Transport (Peyré & Cuturi, 2019). However, permutation alignment alone is insufficient for LoRA, where *invertible basis transformations* $BA = (BT)(T^{-1}A)$ induce $GL(\cdot)$ symmetries that are not combinatorial (Putterman et al., 2024). FedRot-LoRA (Zhang et al., 2026) rotates client factors toward the global ones before averaging, and GLoRA (Chen et al., 2026) projects the average update onto a consensus of the clients' left factors. However, these methods require both factors at the server.

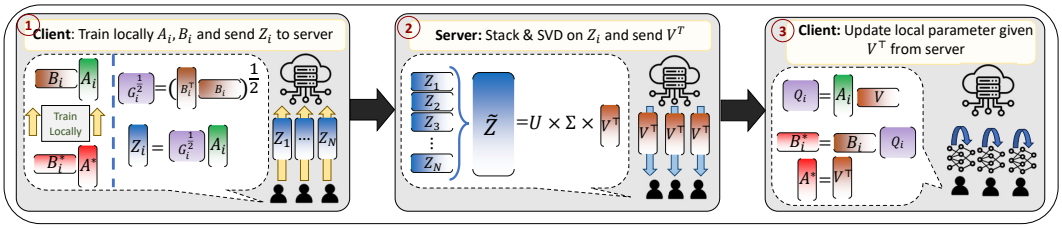


Figure 3: An overview of PERSIA in *PFL*. Z_i and B_i^* are computed at the clients, given A^* from the server; B_i and the classifier head never leave the client.

Full-rank merging reconstructs and averages full-rank updates, $\frac{1}{N}\sum_i B_i A_i$, and then refactors them via SVD (Stoica et al., 2025). When both factors are available, the same truncation can also be computed from thin factorizations of the stacked factors (Appendix D.2). Either way, it requires both factors and is infeasible for FL with the FedSA-LoRA architecture (Guo et al., 2025; Li et al., 2025).

LoRA merging methods, including sign-aware sparsified merging (Yadav et al., 2023), task arithmetic (Ilharco et al., 2023), and LoRA-specific alignment heuristics (Zhao et al., 2025), often improve practical merging, but they (i) do not provide a principled low-rank proxy for full-rank averaging, (ii) require heuristics or data-dependent selection, and (iii) do not extend to PFL where only the right factor is shared. Taken together, existing methods typically sacrifice (i) strict data-free aggregation, (ii) purely low-rank computation, or (iii) compatibility with personalized protocols that share only the LoRA right factor; PERSIA satisfies all three.

3 Methodology

This section presents PERSIA for personalized FL. We formalize the one-sided protocol, define the aggregation error that a shared right factor induces on the personalized client updates, and show that PERSIA minimizes it exactly among all shared rank-r right factors. PERSIA first constructs a common base and then rebases each client solution (Figure 3; algorithms in Appendix A). The global-FL variant, in which the server also averages the rebased left factors, is deferred to Appendix E.

3.1 Personalized Right-Subspace Invariant Aggregation

Notation. For a LoRA-injected linear layer with frozen weight matrix $W \in \mathbb{R}^{d\times k}$, client i learns a rank-r update $\Delta W_i = B_i A_i$ with $B_i \in \mathbb{R}^{d\times r}$ and $A_i \in \mathbb{R}^{r\times k}$, where $r \ll \min\{d, k\}$. For simplicity, we assume $k = d$, while all derivations hold verbatim for general k.

Personalized protocol. In round t, all clients hold the shared right factor A^t and private left factors B_i^t (and private heads). Each client trains both factors for τ local steps and uploads a right-factor-sized message; the server returns a new shared A^{t+1}, after which each client updates its private B_i. A one-sided aggregation rule thus outputs a shared $A' \in \mathbb{R}^{r\times k}$ and client-specific $\{B_i'\}_{i=1}^N$; FedSA-LoRA (Guo et al., 2025) uses $A' = \frac{1}{N}\sum_i A_i$ and $B_i' = B_i$. Aggregation should preserve what each client has learned, which we measure by the personalized aggregation error

$$E_{\mathbf{PFL}}(A', \{B_i'\}_{i=1}^N) := \Big(\frac{1}{N}\sum_{i=1}^N \|\Delta W_i - B_i' A'\|_F^2\Big)^{1/2}. \tag{2}$$

Client-side. In each round, clients train low-rank factors A_i and B_i. To make the low-rank geometry comparable across clients, PERSIA uses the Gram matrix of the local left factor,

$$G_i := (B_i^\top B_i)^{1/2} \in \mathbb{R}^{r\times r}, \qquad Z_i := G_i A_i \in \mathbb{R}^{r\times d},$$

where G_i is the positive semidefinite square root, which exists for every B_i, including rank-deficient ones. When B_i has full column rank, the *polar decomposition* $B_i = O_i G_i$ with $O_i := B_i G_i^{-1}$, $O_i^\top O_i = I_r$, gives the geometric picture of Figure 2: $B_i A_i = O_i(G_i A_i) = O_i Z_i$, so whitening strips the client-specific orthonormal frame O_i from the update. Gram-whitening converts the client-induced metric into a Euclidean one, which is the key step that makes a common right-subspace

objective well-defined across clients:

$$\Delta W_i = B_i A_i \quad \Rightarrow \quad \|\Delta W_i\|_F^2 = \mathrm{tr}(A_i^\top B_i^\top B_i A_i) = \|G_i A_i\|_F^2 = \|Z_i\|_F^2, \tag{3}$$

and for any orthogonal projector $P = P^\top = P^2 \in \mathbb{R}^{d\times d}$,

$$\|\Delta W_i (I - P)\|_F^2 = \|Z_i (I - P)\|_F^2. \tag{4}$$

Both identities follow from $Z_i^\top Z_i = A_i^\top B_i^\top B_i A_i = \Delta W_i^\top \Delta W_i$ and hold for rank-deficient B_i (Lemma F.1); in practice, we compute G_i by an eigendecomposition and may add a small ridge, $(B_i^\top B_i + \varepsilon I_r)^{1/2}$, which adds $\varepsilon A_i^\top A_i$ to $Z_i^\top Z_i$; exact invariance to the factorization holds at $\varepsilon = 0$.

Server-side. The server receives the whitened factors Z_i and forms $\tilde{Z} \in \mathbb{R}^{(Nr)\times d}$ as:

$$\tilde{Z} := \left[Z_1^\top,\ Z_2^\top,\ \ldots,\ Z_N^\top\right]^\top \in \mathbb{R}^{(Nr)\times d}. \tag{5}$$

Afterward, the server computes the top-r right singular vectors $V_r \in \mathbb{R}^{d\times r}$ via low-rank SVD. Equivalently, V_r solves the classical principal subspace problem, where

$$V_r \in \underset{V\in\mathbb{R}^{d\times r}:\ V^\top V = I_r}{\arg\min} \left\|\tilde{Z} - \tilde{Z} V V^\top\right\|_F^2. \tag{6}$$

The low-rank SVD can be approximated via an iterative procedure, which is much cheaper than computing the full-rank SVD and truncating the top-r ranks (Halko et al., 2011). The server broadcasts $A^\star := V_r^\top$ as the new shared right factor.

Client update. After receiving the $V_r^\top$ from the server, each client relocates the B_i from their original A_i subspace to $V_r^\top$ via the transformation matrix $Q_i = A_i V_r \in \mathbb{R}^{r\times r}$ as $B_i^\star = B_i Q_i$. Since $A^\star$ has orthonormal rows, $B_i^\star = \Delta W_i V_r$ is the least-squares solution of $\min_C \|\Delta W_i - C A^\star\|_F$, and $B_i^\star A^\star$ is equivalent to a projection of the full-rank personalized solution into the common space:

$$B_i^\star A^\star = (B_i Q_i) V_r^\top = B_i A_i V_r V_r^\top = (B_i A_i)(V_r V_r^\top). \tag{7}$$

Hence, a client update that lies in the shared subspace is preserved exactly; otherwise it is replaced by its minimum-error projection. **Communication.** Each client uploads $Z_i \in \mathbb{R}^{r\times d}$, the same size as the A_i uploaded by FedSA-LoRA, and neither B_i nor ΔW_i leaves the client, although Z_i reveals $Z_i^\top Z_i = \Delta W_i^\top \Delta W_i$; PERSIA adds no privacy guarantee beyond standard FL. Our implementation equivalently uploads A_i together with the $r \times r$ Gram matrix $B_i^\top B_i$ and forms Z_i at the server, which adds r^2 scalars per layer.

3.2 Optimality and invariance of PERSIA

Based on the above definitions, Proposition 3.1 characterizes the PERSIA error in PFL: the error is exactly captured by the stacked residual tail, and no other shared rank-r right factor can do better. The formal proofs are provided in Appendix F.

Proposition 3.1 (Optimal shared right subspace in PFL)**.** *Let $\tilde{Z}$ and V_r be as in equation 5 and equation 6, with arbitrary B_i and A_i, and define the stacked residual tail $\mathcal{R} := \sum_{j>r} \sigma_j^2(\tilde{Z}) = \min_{V^\top V = I_r} \|\tilde{Z} - \tilde{Z} V V^\top\|_F^2$. Then, for every $A' \in \mathbb{R}^{r\times k}$ and all $\{B_i'\}_{i=1}^N$,*

$$E_{\mathbf{PFL}}(A', \{B_i'\}_{i=1}^N)^2 \ \geq\ \frac{\mathcal{R}}{N} \ =\ E_{\mathbf{PFL}}(A^\star, \{B_i^\star\}_{i=1}^N)^2,$$

where $A^\star = V_r^\top$ and client-specific $B_i^\star = B_i(A_i V_r)$. Moreover, $\tilde{Z}^\top \tilde{Z} = \sum_i \Delta W_i^\top \Delta W_i$, so $\mathcal{R}$ and the selected subspace (up to ties among singular values) depend only on the updates $\{\Delta W_i\}$ and are invariant to every reparameterization $(B_i, A_i) \mapsto (B_i T_i, T_i^{-1} A_i)$ with $T_i \in GL(r)$.

Proposition 3.1 covers every one-sided rule, including factor averaging ($A' = \frac{1}{N}\sum_i A_i$, $B_i' = B_i$), rebasing onto the row space of $\frac{1}{N}\sum_i A_i$, and a truncated SVD of the unwhitened stack $[A_1; \ldots; A_N]$: given the same client updates, each incurs at least the error of PERSIA in every round, and the gap can be arbitrarily large. When all clients learn the same functionality in different coordinates, the error vanishes entirely:

Corollary 3.2 (Eliminating $GL(\cdot)$-symmetry error)**.** *Suppose all clients learn the same functionality under various symmetric representations. $\Delta W_i = W^* = BA = (BT_i)(T_i^{-1}A)$, where $T_i \in \mathbb{R}^{r\times r}$ is a client-specific invertible transformation. More generally, suppose the row spaces of all ΔW_i span a subspace of dimension at most r. Then $\mathcal{R} = 0$ and PERSIA preserves every client update exactly, $E_{\mathbf{PFL}}(A^\star, \{B_i^\star\}_{i=1}^N) = 0$, whereas factor averaging can lose the entire update.*

Scalability. Each client forms its $r \times r$ Gram matrix and rebases B_i in $\mathcal{O}(dr^2)$ time, and the server computes a randomized truncated SVD of the $(Nr)\times d$ stack in $\mathcal{O}(Ndr^2)$ time and $\mathcal{O}(Ndr)$ memory per layer, versus $\mathcal{O}(Ndr)$ for factor averaging; no $d \times d$ matrix is formed. Our experiments use an exact thin SVD, $\mathcal{O}(d(Nr)^2)$, negligible at $N = 10$ (Appendices D.2 and D.3).

4 Convergence analysis

Proposition 3.1 holds for any fixed client updates. We now analyze the optimization that PERSIA actually performs: joint stochastic updates of the LoRA factors. Let $F_i(A, B) := f_i(W_0 + BA)$, where f_i is the objective of client i in the induced weight, and write $x = (A, B)$. Because the algorithm updates the factors, we measure stationarity in the factor variables,

$$\mathcal{G}_i(A,B) := \|\nabla_A F_i(A,B)\|_F^2 + \|\nabla_B F_i(A,B)\|_F^2, \qquad \nabla_A F_i = B^\top \nabla f_i, \quad \nabla_B F_i = \nabla f_i\, A^\top.$$

Each round starts from the broadcast A^t and the private B_i^t; every client runs τ local SGD steps $x_i^{t,s+1} = x_i^{t,s} - \eta\,(B^\top g_i,\ g_i A^\top)$ with a stochastic gradient g_i of f_i, and PERSIA maps the local iterates $\{x_i^{t,\tau}\}$ to $\{x_i^{t+1,0}\}$. We consider one adapted layer for notation; the multi-layer case sums the residuals over layers. All clients participate in every round.

Assumption 4.1 (Smooth)**.** *For each client objective f_i and $W, W' \in \mathbb{R}^{d\times k}$, $\|\nabla f_i(W) - \nabla f_i(W')\|_F \le L\|W - W'\|_F$.*

Assumption 4.2 (Unbiased gradients and bounded variance)**.** *Each client uses stochastic gradients $g_i(W;\xi)$ with $\mathbb{E}[g_i(W;\xi)] = \nabla f_i(W)$ and $\mathbb{E}\|g_i(W;\xi) - \nabla f_i(W)\|_F^2 \le \sigma^2$.*

Assumption 4.3 (Bounded factor iterates)**.** *All local, broadcast, and rebased iterates satisfy $\|A_i^{t,s}\|_2 \le \rho_A$ and $\|B_i^{t,s}\|_2 \le \rho_B$ almost surely.*

Bounded factor iterates are the standard hypothesis in convergence analyses of federated and decentralized LoRA (Chen et al., 2025; Park & Klabjan, 2024; Ghiasvand et al., 2025; Saadati et al., 2026), since the gradient of F_i in the factors is not globally Lipschitz. Ours is a spectral bound; it implies bounded gradients on the bounded set, and no bounded-dissimilarity condition (Wang et al., 2020b) is assumed. Define

$$G_\nabla := \max_i \|\nabla f_i(W_0)\|_F + L\rho_A\rho_B\sqrt{r}, \quad \Gamma^2 := G_\nabla^2 + \sigma^2, \quad \sigma_F^2 := (\rho_A^2 + \rho_B^2)\sigma^2,$$
$$L_F := G_\nabla + L(\rho_A + \rho_B)^2, \quad \Phi := \rho_B^4\Gamma^2, \quad \mathcal{X} := \{(A,B) : \|A\|_2 \le \rho_A,\ \|B\|_2 \le \rho_B\},$$
$$D := \tfrac{1}{N}\textstyle\sum_i \big(F_i(x_i^{0,0}) - \min_{\mathcal{X}} F_i\big).$$

Under the assumptions, G_∇ bounds $\|\nabla f_i\|_F$ on $\mathcal{X}$, F_i is L_F-smooth on $\mathcal{X}$, and the factor-gradient estimates have variance at most σ_F^2 (Lemma G.1). Let $\mathcal{R}^t := \sum_{j>r}\sigma_j^2(\tilde{Z}^t)$ be the stacked residual tail of Proposition 3.1 at the end of round t.

Theorem 4.4 (Convergence of PERSIA-PFL to a factor-stationary neighborhood)**.** *Under Assumptions 4.1 to 4.3, with τ local steps and step size $0 < \eta \le \frac{1}{L_F}$:*

(i) Residual form. *With no bound assumed on the residuals,*

$$\frac{1}{NT\tau}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\sum_{s=0}^{\tau-1}\mathbb{E}\,\mathcal{G}_i(x_i^{t,s}) \ \le\ \frac{2D}{\eta\tau T} + L_F\eta\sigma_F^2 + \frac{2}{\eta\tau T}\sum_{t=0}^{T-1}\Big(G_\nabla\sqrt{\bar{\mathcal{R}}_t} + \frac{L}{2}\bar{\mathcal{R}}_t\Big), \quad \bar{\mathcal{R}}_t := \frac{\mathbb{E}[\mathcal{R}^t]}{N}. \tag{8}$$

(ii) Rate. *Moreover, $\bar{\mathcal{R}}_t \le \Phi\eta^2\tau^2$. With $D > 0$, $H := L_F\sigma_F^2 + L\Phi\tau > 0$, and $\eta = \min\{1/L_F, \sqrt{2D/(\tau T H)}\}$,*

$$\frac{1}{NT\tau}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\sum_{s=0}^{\tau-1}\mathbb{E}\,\mathcal{G}_i(x_i^{t,s}) \ \le\ \underbrace{2G_\nabla\sqrt{\Phi}}_{\textit{floor}} + 2\sqrt{\frac{2DH}{\tau T}} \ +\ \frac{2DL_F}{\tau T}. \tag{9}$$

Table 1: Personalized accuracy (%) on GLUE with RoBERTa-Large (10 clients, Dirichlet $\alpha = 0.5$). **Top, middle:** final accuracy, standard and communication-sparse schedules. **Bottom:** the sparse runs right after their first aggregation (local only: the clients' models after the same local phase).

Method	Uplink (MiB)	MNLI	QNLI	QQP	SST-2	RTE	Avg.
Standard schedule ($\tau = 40$, $T = 80$): final accuracy							
Local only	0	$90.00_{\pm 0.23}$	$91.04_{\pm 1.19}$	$88.51_{\pm 0.30}$	$96.14_{\pm 0.22}$	$85.09_{\pm 0.44}$	91.42
FFA-LoRA, personal head (Sun et al., 2024)	0.75	$85.67_{\pm 0.41}$	$87.05_{\pm 0.61}$	$85.46_{\pm 0.39}$	$95.56_{\pm 0.36}$	$85.19_{\pm 0.51}$	88.44
FedProx, personal head (Li et al., 2020)	1.50	$90.07_{\pm 0.35}$	$92.02_{\pm 0.26}$	$88.91_{\pm 0.21}$	$96.11_{\pm 0.23}$	$85.67_{\pm 1.24}$	91.78
FedAvg-LoRA, personal head (McMahan et al., 2017)	1.50	$\mathbf{90.58}_{\pm \mathbf{0.25}}$	$\mathbf{93.82}_{\pm \mathbf{0.09}}$	$\mathbf{89.73}_{\pm \mathbf{0.23}}$	$\mathbf{96.45}_{\pm \mathbf{0.20}}$	$85.47_{\pm 0.52}$	**92.65**
FedEx-LoRA, personal head (Singhal et al., 2025)	1.50	$\mathbf{90.58}_{\pm \mathbf{0.33}}$	$93.62_{\pm 0.51}$	$89.70_{\pm 0.20}$	$96.35_{\pm 0.10}$	$\mathbf{86.39}_{\pm \mathbf{1.10}}$	92.56
FedSA-LoRA (Guo et al., 2025)	0.75	$90.24_{\pm 0.13}$	$91.72_{\pm 0.53}$	$89.10_{\pm 0.08}$	$96.16_{\pm 0.29}$	$85.39_{\pm 0.81}$	91.80
PERSIA (ours)	0.76	$89.69_{\pm 0.66}$	$92.39_{\pm 0.15}$	$88.97_{\pm 0.01}$	$95.95_{\pm 0.22}$	$84.53_{\pm 0.69}$	91.75
Communication-sparse schedule ($\tau = 1280$, $T = 3$): final accuracy							
Local only	0	$88.94_{\pm 1.38}$	$91.64_{\pm 0.40}$	$88.98_{\pm 0.01}$	$96.10_{\pm 0.25}$	$84.88_{\pm 0.16}$	91.42
FedAvg-LoRA, personal head (McMahan et al., 2017)	1.50	$89.86_{\pm 0.31}$	$\mathbf{93.05}_{\pm \mathbf{0.31}}$	$88.67_{\pm 0.37}$	$96.30_{\pm 0.31}$	$84.46_{\pm 0.43}$	**91.97**
FedSA-LoRA (Guo et al., 2025)	0.75	$\mathbf{90.14}_{\pm \mathbf{0.20}}$	$92.12_{\pm 0.25}$	$89.08_{\pm 0.19}$	$\mathbf{96.32}_{\pm \mathbf{0.09}}$	$83.66_{\pm 0.54}$	91.92
PERSIA (ours)	0.76	$90.13_{\pm 0.30}$	$92.14_{\pm 0.20}$	$\mathbf{89.35}_{\pm \mathbf{0.28}}$	$96.11_{\pm 0.05}$	$\mathbf{85.36}_{\pm \mathbf{0.31}}$	91.93
Same runs, right after the first *aggregation ($\tau = 1280$, one round)*							
Local only (no aggregation)	0	$88.63_{\pm 1.42}$	$90.87_{\pm 0.60}$	$\mathbf{87.99}_{\pm \mathbf{0.13}}$	$95.79_{\pm 0.27}$	$\mathbf{86.51}_{\pm \mathbf{0.81}}$	90.82
FedAvg-LoRA, personal head (McMahan et al., 2017)	1.50	$65.88_{\pm 4.33}$	$77.10_{\pm 4.15}$	$80.39_{\pm 1.93}$	$94.75_{\pm 0.73}$	$85.26_{\pm 0.34}$	79.53
FedSA-LoRA (Guo et al., 2025)	0.75	$80.27_{\pm 4.13}$	$86.99_{\pm 2.49}$	$84.17_{\pm 1.19}$	$95.28_{\pm 0.58}$	$84.02_{\pm 0.48}$	86.68
PERSIA (ours)	0.76	$\mathbf{89.28}_{\pm \mathbf{0.16}}$	$\mathbf{91.13}_{\pm \mathbf{0.43}}$	$87.93_{\pm 0.06}$	$\mathbf{96.02}_{\pm \mathbf{0.23}}$	$85.92_{\pm 0.66}$	**91.09**

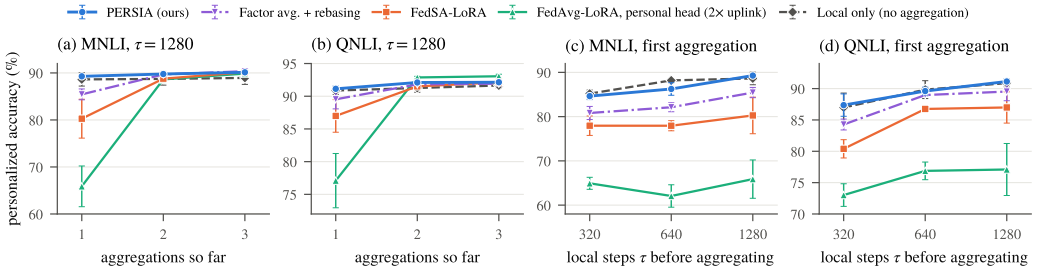


Figure 4: What each aggregation preserves (mean ± sd, 3 seeds). **(a, b)** Accuracy after each of the three aggregations at $\tau = 1280$ (bottom and middle blocks of Table 1). **(c, d)** Accuracy right after the first aggregation (Table 9). Local only (dashed): the clients' own models after training.

Discussion. Part (i) makes no residual assumption: aggregation enters only through the SVD tail $\bar{\mathcal{R}}_t$. The same bound holds for any one-sided rule with its own mean squared displacement, which by Proposition 3.1 is at least $\mathcal{R}^t/N$ from the same pre-aggregation state; hence PERSIA minimizes this term each round. Part (ii) does not yield a better asymptotic rate than factor averaging, since both induce $\mathcal{O}(\eta\tau)$ displacement. Its floor $2G_\nabla \rho_B^2 \Gamma$ is informative when $\rho_A^2 > \rho_B^2(2\Gamma/G_\nabla - 1)$, as is typical when B starts at zero and the broadcast A has orthonormal rows. With the stated step size, the floor is independent of τ, and if the residual vanishes every round (Corollary 3.2), part (i) recovers the floor-free $\mathcal{O}(1/\sqrt{\tau T})$ rate. The result concerns factor stationarity under full participation and exact SVD, excluding the ridge, and adaptive optimizers used experimentally (Appendix G).

5 Experimental Setup

We evaluate PERSIA in the personalized protocol of Section 3 on **NLU classification** on the GLUE suite (Wang et al., 2019) (MNLI, SST-2, QNLI, QQP, RTE) using RoBERTa-Large (355M parameters, encoder-only) (Liu et al., 2019). The left factors B_i and the classification heads remain on the clients in every personalized method. A two-factor global protocol, including generative reasoning on GSM8K (Cobbe et al., 2021) with Qwen2.5-3B-Instruct (Qwen Team, 2024), is reported in Appendix E. All implementations are built on FederatedScope-LLM (Kuang et al., 2024).

Data Partitioning. We distribute the training data across 10 clients with a Dirichlet distribution over class labels with concentration $\alpha = 0.5$ (Hsu et al., 2019). Variations of label skew, client count, rank, participation, and compute are reported in Table 3 and Appendix C.

Evaluation. Each client holds out part of its local data, and we report the unweighted mean of client accuracies (Collins et al., 2021; Li et al., 2021). Clients are evaluated right after an aggregation, once

Table 2: Long local horizons: what rebasing and whitening each contribute. **Left:** $\tau = 1280$, $T = 3$ (3,840 local steps). **Right:** matched compute, $\tau = 1067$, $T = 3$ (3,201 steps).

	$\tau = 1280, T = 3$				Matched compute: $\tau = 1067, T = 3$			
	MNLI		QNLI		MNLI		QNLI	
Method	1st	final	1st	final	1st	final	1st	final
Local only (no aggregation)	$88.63_{\pm 1.42}$	$88.94_{\pm 1.38}$	$90.87_{\pm 0.60}$	$91.64_{\pm 0.40}$	$88.95_{\pm 0.13}$	$89.63_{\pm 0.30}$	$\mathbf{91.17}_{\pm \mathbf{0.52}}$	$91.73_{\pm 0.49}$
FedAvg-LoRA, personal head	$65.88_{\pm 4.33}$	$89.86_{\pm 0.31}$	$77.10_{\pm 4.15}$	$\mathbf{93.05}_{\pm \mathbf{0.31}}$	$64.24_{\pm 0.10}$	$89.63_{\pm 0.26}$	$76.19_{\pm 4.01}$	$\mathbf{93.16}_{\pm \mathbf{0.21}}$
FedSA-LoRA	$80.27_{\pm 4.13}$	$90.14_{\pm 0.20}$	$86.99_{\pm 2.49}$	$92.12_{\pm 0.25}$	$82.91_{\pm 1.25}$	$90.32_{\pm 0.07}$	$86.90_{\pm 1.55}$	$91.75_{\pm 0.25}$
Factor avg. + rebasing	$85.44_{\pm 1.16}$	$\mathbf{90.37}_{\pm \mathbf{0.17}}$	$89.53_{\pm 1.46}$	$91.92_{\pm 0.33}$	$86.51_{\pm 0.84}$	$90.15_{\pm 0.04}$	$88.61_{\pm 1.27}$	$91.96_{\pm 0.33}$
PERSIA (ours)	$\mathbf{89.28}_{\pm \mathbf{0.16}}$	$90.13_{\pm 0.30}$	$\mathbf{91.13}_{\pm \mathbf{0.43}}$	$92.14_{\pm 0.20}$	$\mathbf{89.28}_{\pm \mathbf{0.26}}$	$\mathbf{90.74}_{\pm \mathbf{0.40}}$	$89.99_{\pm 0.97}$	$91.62_{\pm 0.25}$

Table 3: Robustness at $\tau = 40$, $T = 80$ (final accuracy, %, mean$_{\pm \text{sd}}$ over 3 seeds): label skew (Dirichlet α, IID), $N = 20$ clients, rank r, and partial participation (3 of 10 clients per round). Further settings are in Appendix C.4 and Tables 10 and 11.

Method	$\alpha = 0.1$	$\alpha = 1.0$	IID	$N = 20$	$r = 4$	$r = 32$	3 of 10
MNLI							
FedAvg-LoRA, personal head (2× uplink)	$97.08_{\pm 0.31}$	$90.82_{\pm 0.05}$	$88.42_{\pm 0.10}$	$92.07_{\pm 0.12}$	$90.31_{\pm 0.19}$	$91.31_{\pm 0.14}$	$90.02_{\pm 0.31}$
FedSA-LoRA	$97.44_{\pm 0.13}$	$90.37_{\pm 0.16}$	$87.41_{\pm 0.10}$	$90.59_{\pm 0.07}$	$72.65_{\pm 11.27}$	$88.47_{\pm 1.02}$	$87.44_{\pm 1.90}$
PERSIA (ours)	$97.15_{\pm 0.23}$	$90.13_{\pm 0.11}$	$86.98_{\pm 0.06}$	$90.79_{\pm 0.13}$	$73.76_{\pm 11.65}$	$89.32_{\pm 0.03}$	$88.05_{\pm 0.64}$
QNLI							
FedAvg-LoRA, personal head (2× uplink)	$92.54_{\pm 0.68}$	$95.62_{\pm 0.15}$	$93.63_{\pm 0.24}$	$94.09_{\pm 0.13}$	$93.37_{\pm 0.31}$	$94.57_{\pm 0.22}$	$93.69_{\pm 0.12}$
FedSA-LoRA	$93.08_{\pm 0.25}$	$95.21_{\pm 0.12}$	$92.33_{\pm 0.22}$	$93.18_{\pm 0.19}$	$88.44_{\pm 2.38}$	$92.43_{\pm 0.24}$	$91.41_{\pm 0.70}$
PERSIA (ours)	$92.77_{\pm 0.34}$	$95.52_{\pm 0.05}$	$92.04_{\pm 0.07}$	$93.16_{\pm 0.13}$	$87.71_{\pm 2.78}$	$91.32_{\pm 0.32}$	$91.19_{\pm 0.16}$

they have merged the broadcast factors and rebased B_i, and before any further local training. A run's score is its final-round accuracy; for the communication-sparse runs we also report the accuracy after the first aggregation. We give the mean $\pm$ population standard deviation over three seeds and the per-client majority floor (Appendix C). On RTE, four of the ten clients have no held-out examples, and the majority floor (88.1%) exceeds that of every method, so we draw no conclusions from RTE and exclude it from the averages.

Communication budget. We fix the local work at $\tau T = 3{,}200$ steps per client and vary $\tau \in \{10, 20, \ldots, 640\}$ ($T = 3200/\tau$). The communication-sparse setting uses $\tau = 1280$ and $T = 3$ (3,840 steps), and the standard schedule $\tau = 40$ and $T = 80$.

Baselines. All baselines keep the classification head local. *Local only* trains each client in isolation. *FedSA-LoRA* (Guo et al., 2025) averages the A_i and keeps B_i local; it is the factor-averaging rule that PERSIA replaces, at the same uplink. *FedAvg-LoRA* (McMahan et al., 2017), *FedEx-LoRA* (Singhal et al., 2025), and *FedProx* (Li et al., 2020) share the full adapter with a personalized head (Arivazhagan et al., 2019; Collins et al., 2021) (2× the uplink), and *FFA-LoRA* (Sun et al., 2024) freezes A and shares B. Methods that need B_i at the server are compared in the global protocol (Appendix E). We also report PERSIA without whitening (SVD of the stacked A_i) and factor averaging followed by PERSIA's rebasing.

Hyperparameters. All methods share AdamW (Loshchilov & Hutter, 2019) with learning rate 2×10^{-4} and weight decay 0.01, LoRA on the query and value projections with $r = 8$, $\alpha = 16$, and dropout 0.05, batch size 16, maximum sequence length 128, and bf16 mixed precision; the optimizer state is re-initialized at every local phase. PERSIA uses a Gram square root with a small ridge and an exact SVD, plus three additions not covered by Proposition 3.1 (Appendix B.3): an error-feedback buffer for the truncation residual (Stich et al., 2018; Karimireddy et al., 2019), random exploration of the last two shared directions, and a guard that resets B_i after a collapsed local phase. Removing the buffer ($\tau = 40$) or exploration and the guard ($\tau = 1280$) changes accuracy by at most 0.5 points (Appendix B). The buffer is empty at the first aggregation, so at that point the runs without exploration and the guard (Table 5) apply the rule of Proposition 3.1 up to the ridge.

6 Results and Discussion

Right after aggregating, PERSIA matches local training. The bottom block of Table 1 evaluates the communication-sparse runs right after their first aggregation. With PERSIA, the aggregated

personalized models are as accurate as the clients' local models (91.09 versus 90.82 on average), whereas factor averaging loses 4.14 points and two-factor averaging with personalized heads 11.29; on MNLI, the losses are 8.36 and 22.75 points. Across $\tau \in \{320, 640, 1280\}$ (Figure 4(c,d)), PERSIA stays within 1.94 points of local training on MNLI and QNLI, while factor averaging loses 3.09 to 10.25 points and two-factor averaging 12.96 to 26.12, consistent with Proposition 3.1.

Further rounds repair the other rules, so final accuracy is on par. Local training absorbs most of the other rules' aggregation error within one round (Figure 4(a,b)): after the second aggregation, factor averaging and two-factor averaging are back at the level of local training on MNLI (88.78 and 88.63 versus 88.73), one point below PERSIA (89.77). After the third, FedAvg-LoRA (91.97), PERSIA (91.93), and FedSA-LoRA (91.92) are within 0.05 points on average and above local training (91.42), with PERSIA uploading half as much as FedAvg-LoRA. The advantage of PERSIA thus lies in what each aggregation preserves, which matters when only one or two aggregations are possible or models are used right after aggregating, not in accuracy after several rounds.

Frequent communication. At $\tau = 40$, all methods except FFA-LoRA lie within 1.3 points on average; full-adapter sharing (FedAvg-/FedEx-LoRA) is the most accurate at twice the uplink, and PERSIA is on par with FedSA-LoRA (91.75 versus 91.80). Frequent synchronization keeps the residual small (Theorem 4.4); at $\tau \leq 20$, PERSIA trails factor averaging on MNLI (Table 7).

Whitening ablation and communication. Removing Gram whitening, i.e., selecting the subspace from the stacked raw A_i, degrades PERSIA by 24.85 points on MNLI and 16.67 points on QNLI (Table 4), bringing it to the majority-class floor. The SVD then ranks directions by factor magnitude instead of update energy.

Table 4: Ablations at $\tau = 40$ (accuracy).

Variant	MNLI	QNLI
PERSIA	$89.69_{\pm 0.66}$	$92.39_{\pm 0.15}$
w/o Gram whitening	$64.84_{\pm 0.91}$	$75.72_{\pm 0.09}$
w/o error feedback	$90.13_{\pm 0.26}$	$92.12_{\pm 0.36}$
Factor avg. + rebasing	$90.27_{\pm 0.03}$	$92.45_{\pm 0.57}$
Majority-class floor	68.70	75.50

Whitening and rebasing at long horizons. Table 2 separates the two steps of PERSIA. At the first aggregation with $\tau = 1280$, rebasing onto the averaged factor recovers 5.17 of the 8.36 points that factor averaging loses on MNLI (2.54 of 3.88 on QNLI), and PERSIA's whitened subspace recovers the rest; across $\tau \in \{320, 640, 1280\}$, whitening adds 3.81 to 4.13 points on MNLI and 0.64 to 3.10 on QNLI on top of rebasing (Table 9). At matched compute ($\tau = 1067$, 3,201 local steps) the picture is the same; at the final round, PERSIA is the most accurate on MNLI and two-factor sharing on QNLI, at twice the uplink.

Heterogeneous tasks and robustness. With one of six GLUE tasks per pair of 12 clients, PERSIA reaches $85.79_{\pm 0.50}$ versus $84.93_{\pm 0.16}$ for FedSA-LoRA. Across label skew, client count, rank, and partial participation (Table 3), PERSIA stays within 1.2 points of FedSA-LoRA, and two-factor sharing is the most accurate, at twice the uplink, except at $\alpha = 0.1$. When clients run very different numbers of local steps, clients with larger updates dominate the energy-weighted subspace of Proposition 3.1 and PERSIA trails factor averaging (QNLI, severe: 88.29 versus 91.33, Table 11); reweighting the stacked blocks (Remark F.6) is a natural remedy.

Global protocol. When both factors are shared (Appendix E), PERSIA is the most accurate of the compared methods on MNLI and QNLI at $\tau \in \{40, 640\}$. On GSM8K with Qwen2.5-3B-Instruct, it trails dense average-then-SVD and FLoRA.

7 Conclusion

For personalized federated LoRA, PERSIA whitens, recovers a shared right subspace by SVD, and rebases each private left factor; its core rule is exactly optimal among shared rank-r right factors and satisfies a factor-stationarity bound whose floor does not grow with the local horizon. Empirically, right after aggregation, PERSIA's personalized models are as accurate as the clients' local models, whereas factor averaging and two-factor averaging lose up to 10 and 26 points, respectively. Further local training repairs the losses of the other rules, and final accuracy is on par with half the uplink of two-factor sharing. PERSIA is therefore most useful when communication is sparse or when models are used immediately after aggregation.

AI USE STATEMENT

We used generative AI tools to edit and revise the manuscript text, check bibliographic metadata against primary sources, and organize result tables and figures from our experiment logs. We did not use generative AI tools to generate experimental results. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 5 and Appendix B specify the data partitioning, evaluation protocol, schedules, and hyperparameters; Appendix A gives the algorithms; Appendices C and E report every configuration we ran with per-seed statistics; and Appendices F and G contain complete proofs of all theoretical statements, with every assumption stated explicitly. Anonymous code with all baselines and configurations is provided as supplementary material.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. LoRA-FAIR: Federated LoRA fine-tuning with aggregation and initialization refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3737–3746, 2025.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. FedALT: Federated fine-tuning through adaptive local training with rest-of-world LoRA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19728–19736, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Jinqian Chen, Chang Liu, and Jihua Zhu. Beyond factor aggregation: Gauge-aware low-rank server representations for federated LoRA. *arXiv preprint arXiv:2605.06733*, 2026.

Xin Chen, Shuaijun Chen, Omid Tavallaie, Nguyen Tran, Shuhuang Xiang, and Albert Zomaya. Convergence analysis of aggregation-broadcast in LoRA-enabled distributed fine-tuning. *arXiv preprint arXiv:2508.01348*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2089–2099. PMLR, 2021.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with Moreau envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21394–21405, 2020.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Wenzhi Fang, Dong-Jun Han, Liangqi Yuan, Seyyedali Hosseinalipour, and Christopher G. Brinton. Federated sketching LoRA: A flexible framework for heterogeneous collaborative fine-tuning of LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Sajjad Ghiasvand, Mahnoosh Alizadeh, and Ramtin Pedarsani. Decentralized low-rank fine-tuning of large language models. *arXiv preprint arXiv:2501.15361*, 2025.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *International Conference on Learning Representations*, 2025.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Jie Hao, Yuman Wu, Ali Payani, Myungjin Lee, and Mingrui Liu. Personalized federated fine-tuning for heterogeneous data: An automatic rank learning approach via two-level LoRA. *arXiv preprint arXiv:2503.03920*, 2025.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261. PMLR, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. FederatedScope-LLM: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5260–5271. ACM, 2024. doi: 10.1145/3637528.3671573.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, pp. 28541–28564, 2023.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6357–6368. PMLR, 2021.

Yongle Li, Bo Liu, Sheng Huang, Zheng Zhang, Xiao-Tong Yuan, and Richang Hong. Communication-efficient and personalized federated foundation model fine-tuning via tri-matrix adaptation. *arXiv preprint arXiv:2503.23869*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yao Lu, Jiaxing Qi, Zhongzhi Luan, Shaohan Huang, Carol Fung, Hailong Yang, and Depei Qian. FDLoRA: Personalized federated learning of large language model via dual LoRA tuning. *arXiv preprint arXiv:2406.07925*, 2024.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. A survey on LoRA of large language models. *Frontiers of Computer Science*, 19(7):197605, 2025. doi: 10.1007/s11704-024-40663-9.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960. doi: 10.1093/qmath/11.1.50.

Haemin Park and Diego Klabjan. Communication-efficient federated low-rank update algorithm and its connection to implicit regularization. *arXiv preprint arXiv:2409.12371*, 2024.

Zihao Peng, Nan Zou, Jiandian Zeng, Guo Li, Ke Chen, Boyuan Li, and Tian Wang. HiLoRA: Hierarchical low-rank adaptation for personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31746–31757, 2026.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073.

Theo Putterman, Derek Lim, Yoav Gelberg, Stefanie Jegelka, and Haggai Maron. Learning on LoRAs: GL-equivariant processing of low-rank weight spaces for large finetuned models. *arXiv preprint arXiv:2410.04207*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Arian Raje, Baris Askin, Divyansh Jhunjhunwala, and Gauri Joshi. Ravan: Multi-head low-rank adaptation for federated fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 48564–48592, 2025.

Hariharan Ramesh and Jyotikrishna Dass. FLoRIST: Singular value thresholding for efficient and accurate federated fine-tuning of large language models. In *Proceedings of Machine Learning and Systems*, volume 8, 2026.

Nastaran Saadati, Zhanhong Jiang, Joshua R. Waite, Shreyan Ganguly, Aditya Balu, Chinmay Hegde, and Soumik Sarkar. DeCAF: Decentralized consensus-and-factorization for low-rank adaptation of foundation models. *Neural Networks*, 202:108992, 2026. doi: 10.1016/j.neunet.2026.108992.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1316–1336. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.67.

Raghav Singhal, Kaustubh Ponkshe, Rohit Vartak, Lav R. Varshney, and Praneeth Vepakomma. Fed-SB: A silver bullet for extreme communication efficiency and performance in (private) federated LoRA fine-tuning. *Transactions on Machine Learning Research*, 2026.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the Knots. In *International Conference on Learning Representations*, 2025.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *International Conference on Learning Representations*, 2024.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603, 2023. doi: 10.1109/TNNLS.2022.3160699.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020a.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623, 2020b.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. FLoRA: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22513–22533, 2024.

Lingling Xu, Haoran Xie, S. Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6107–6126, 2026. doi: 10.1109/TPAMI.2026.3657354.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115, 2023.

Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. Dual-personalizing adapter for federated foundation models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39409–39433, 2024.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. FedRot-LoRA: Mitigating rotational misalignment in federated LoRA. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Ziyu Zhao, Tao Shen, Didi Zhu, Zexi Li, Jing Su, Xuwu Wang, and Fei Wu. Merging LoRAs like playing LEGO: Pushing the modularity of LoRA to extremes through rank-wise clustering. In *International Conference on Learning Representations*, 2025.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62369–62385. PMLR, 2024.

ORGANIZATION OF THE APPENDIX

- **Appendix A** gives the pseudocode of PERSIA for personalized and global FL.
- **Appendix B** specifies training, evaluation, baselines, the implementation of PERSIA, and the compute budget.
- **Appendix C** reports additional results in the personalized protocol.
- **Appendix D** quantifies communication, server complexity, and wall-clock overhead.
- **Appendix E** treats PERSIA in global FL: aggregation rule, error bounds, and experiments.
- **Appendices F and G** contain the proofs of all propositions and theorems.

Conventions. Unless stated otherwise, experiments use the personalized protocol of Section 5: RoBERTa-Large, LoRA rank 8, 10 clients with Dirichlet label skew $\alpha = 0.5$, private left factors B_i and classification heads. Entries are personalized accuracy (%) at the final round, reported as mean$_{\pm\text{sd}}$ with the population standard deviation over three seeds. Bold marks the best entry per column, averages exclude RTE (Section 5), and ‡PERSIA-R denotes the implementation variant of Appendix B.2.

A ALGORITHMS

Algorithm 1 is the personalized procedure evaluated in the main text. Algorithm 2 is the global variant of Appendix E, in which both factors reach the server.

Algorithm 1 PERSIA for personalized FL

Input: N clients, T rounds, LoRA rank r, datasets $\{D_i\}_{i=1}^N$, participating sets $S_t \subseteq \{1,\dots,N\}$ ($S_t = \{1,\dots,N\}$ under full participation)
Init: shared $A \leftarrow \mathcal{N}(0,\sigma)_{r\times d}$; for every client $A_i \leftarrow A$, $B_i \leftarrow 0_{d\times r}$, and a private head h_i
for $t = 1,\dots,T$ **do**
  Client side: every client $i \in S_t$ trains $A_i, B_i, h_i \leftarrow \mathrm{Train}(D_i, B_i, A_i, h_i)$ for τ local steps
  Compute $G_i \leftarrow (B_i^\top B_i + \epsilon I_r)^{1/2}$ and send $Z_i \leftarrow G_i A_i$ to the server
  Server side: stack $\{Z_i\}_{i\in S_t}$ into $\tilde{Z} \in \mathbb{R}^{|S_t|r\times d}$ and compute $(U,\Sigma,V_r^\top) \leftarrow \mathrm{TruncatedSVD}(\tilde{Z}, r)$
  Broadcast $V_r \in \mathbb{R}^{d\times r}$ to all N clients
  Client side: every client, including those not in S_t, sets $Q_i \leftarrow A_i V_r$, $B_i \leftarrow B_i Q_i$, and $A_i \leftarrow V_r^\top$; B_i and h_i stay local
end for

Algorithm 2 PERSIA for global FL

Input: N clients, T rounds, LoRA rank r, datasets $\{D_i\}_{i=1}^N$
Init: $A_i \leftarrow \mathcal{N}(0,\sigma)_{r\times d}$, $\quad B_i \leftarrow 0_{d\times r}$
for $t = 1,\dots,T$ **do**
  Client side: every client trains $A_i, B_i \leftarrow \mathrm{Train}(D_i, B_i, A_i)$
  Send A_i and B_i to the server
  Server side: compute $G_i \leftarrow (B_i^\top B_i + \epsilon I_r)^{1/2}$ and $Z_i \leftarrow G_i A_i$
  Stack Z_i into $\tilde{Z} \in \mathbb{R}^{Nr\times d}$ and compute $(U,\Sigma,V_r^\top) \leftarrow \mathrm{TruncatedSVD}(\tilde{Z}, r)$
  for $i = 1,\dots,N$ **do**
    $Q_i \leftarrow A_i V_r$ and $B_i' \leftarrow B_i Q_i$
  end for
  $\bar{B} \leftarrow \frac{1}{N}\sum_i B_i'$
  Broadcast $(\bar{B}, V_r^\top)$ as the next LoRA factors
end for

B Experimental Details

B.1 Training and evaluation

Training. All methods share one configuration: RoBERTa-Large with LoRA on the query and value projections of all 24 layers (786,432 adapter parameters and a classification head of about 1.05M parameters), rank $r = 8$, scaling $\alpha = 16$, and dropout 0.05; AdamW (Loshchilov & Hutter, 2019) with learning rate 2×10^{-4} and weight decay 0.01; batch size 16; maximum sequence length 128; bf16 autocast with FP32 master weights; and seeds $\{0, 1, 2\}$. The optimizer state is re-initialized at every local phase for every method, including local only, so local-only accuracy varies slightly with τ at a fixed step budget. No method-specific tuning is performed. Clients are partitioned by a Dirichlet distribution over labels ($\alpha = 0.5$ unless stated otherwise) with the FederatedScope-LLM splitter (Kuang et al., 2024).

Evaluation. Each client holds out part of its local data, and the personalized score is the unweighted mean of client accuracies. A model is evaluated right after an aggregation: each client merges the broadcast factors, rebases B_i where applicable, and evaluates before any further local training. Runs are evaluated every $\max\{1, \lfloor T/20 \rfloor\}$ rounds and after the final round; for $T \leq 20$ every round is evaluated, so the first evaluation measures the model right after the first aggregation.

B.2 Baselines

Every personalized method keeps the classification head local. FedSA-LoRA averages the A_i and keeps B_i local. FedAvg-LoRA, FedEx-LoRA, and FedProx share both factors, FedEx-LoRA with its exact residual correction and FedProx with a proximal term; FFA-LoRA freezes A and shares B. Two controls isolate the components of PERSIA:

- **Factor averaging with rebasing** replaces the shared factor by the average $\bar{A}$ of the A_i and rebases $B_i \leftarrow B_i A_i \bar{A}^+$; it differs from FedSA-LoRA only in this client-side step.
- **PERSIA without whitening** applies the same SVD and rebasing to the unwhitened stack $[A_1; \ldots; A_N]$.

PERSIA-R‡ additionally folds the part of each client's update that rebasing discards into the frozen base weights, so no function is lost, at no extra communication. It tracks PERSIA closely and is reported for completeness.

B.3 Implementation of PERSIA

The server forms $Z_i = (B_i^\top B_i + \varepsilon I_r)^{1/2} A_i$ from the uploaded A_i and the $r \times r$ Gram matrix (symmetrized before the square root), stacks the Z_i, and broadcasts the top-r right singular vectors of the stack, computed by an exact thin SVD. The personalized runs add three components that Proposition 3.1 and Algorithm 1 do not cover; the global-protocol runs use only the first.

- **Error feedback.** The residual of the rank-r truncation is stored and appended to the next round's stack as additional rows (Stich et al., 2018; Karimireddy et al., 2019); it requires no B_i.
- **Exploration.** The last two of the $r = 8$ shared directions are replaced by random orthonormal directions orthogonal to the first $r - 2$, drawn with the run's seed (one direction at $r = 4$ and two at $r = 32$).
- **Rebase guard.** Rebasing keeps each client's update up to its projection onto the shared subspace (PERSIA-R keeps it exactly). It therefore also carries over the state of a client whose local phase collapsed to the majority class, which we observed on small skewed shards. A client is flagged when its training accuracy over the last quarter of its local phase is at most two points above its majority-class rate; clients whose majority class exceeds 80% are never flagged. A flagged client sets $B_i = 0$ on the new shared basis instead of rebasing and cannot be flagged again for three of its rounds. The guard never fires at $\tau \geq 1067$; per run, summed over clients and rounds, it fires 3–8 times at $\tau = 80$, 6–12 at $\tau = 40$, 18–50 at $\tau = 20$, and 117–208 at $\tau = 10$.

These components have little effect on accuracy. Removing error feedback changes final accuracy at $\tau = 40$ by at most 0.44 points (Table 4). Removing exploration and the guard changes every entry

of Table 5 by at most 0.36 points, within the seed spread, and PERSIA without them is at least as accurate right after the first aggregation; for PERSIA-R, the changes are at most 0.32 points.

Table 5: Exploration and the rebase guard in PERSIA at $\tau = 1280$, $T = 3$: accuracy right after the first aggregation and at the final round.

	MNLI		QNLI	
Configuration	1st agg.	final	1st agg.	final
Exploration and guard (reported)	$89.28_{\pm 0.16}$	$90.13_{\pm 0.30}$	$91.13_{\pm 0.43}$	$92.14_{\pm 0.20}$
Guard only	$89.38_{\pm 0.29}$	$90.33_{\pm 0.26}$	$91.49_{\pm 0.34}$	$91.91_{\pm 0.72}$
Neither	$89.33_{\pm 0.22}$	$90.43_{\pm 0.30}$	$91.48_{\pm 0.41}$	$91.89_{\pm 0.33}$

B.4 Compute budget

All runs used a single NVIDIA RTX 3090 GPU (24 GB) per worker, with clients simulated by the federated training framework. The reported results come from $1{,}047$ completed runs (951 RoBERTa-Large/GLUE and 96 Qwen2.5-3B/GSM8K runs, about $1{,}080$ GPU-hours), which include a few diagnostic and replication runs that do not enter the tables, and from 82 further runs (about 79 GPU-hours) for the matched-compute, long-horizon rebasing, and exploration and guard controls.

C Additional Results in the Personalized Protocol

This appendix complements Section 6 in four parts: controls at the standard schedule, the number of local steps per round, accuracy right after aggregation, and robustness. The matched-compute comparison ($\tau = 1067$) is reported in Table 2.

C.1 Controls at the standard schedule

Table 6 complements Table 1 ($\tau = 40$, $T = 80$) with the per-client majority-class floor on the same holdouts, factor averaging with rebasing, and PERSIA-R. Every method exceeds the majority floor by a wide margin on MNLI, QNLI, QQP, and SST-2; on RTE, the floor exceeds every method (Section 5).

Table 6: Controls at the standard schedule, $\tau = 40$, $T = 80$ (complements Table 1).

Method	MNLI	QNLI	QQP	SST-2	RTE	Avg.
Majority-class floor (local holdouts)	68.70	75.50	71.50	74.10	88.07	72.45
Factor avg. + rebasing	$90.27_{\pm 0.03}$	$92.45_{\pm 0.57}$	$88.94_{\pm 0.08}$	$96.35_{\pm 0.34}$	$85.96_{\pm 0.62}$	92.00
PERSIA-R‡	$89.78_{\pm 0.18}$	$91.81_{\pm 0.27}$	$88.50_{\pm 0.24}$	$95.02_{\pm 0.15}$	$84.78_{\pm 0.33}$	91.28
PERSIA	$89.69_{\pm 0.66}$	$92.39_{\pm 0.15}$	$88.97_{\pm 0.01}$	$95.95_{\pm 0.22}$	$84.53_{\pm 0.69}$	91.75

C.2 Number of local steps per round

Table 7 fixes the local work at $\tau T = 3{,}200$ steps ($3{,}840$ at $\tau = 1280$) and varies the number of local steps τ between aggregations. On MNLI, the three federated rules lie within 0.52 points of each other for $\tau \geq 80$, while PERSIA trails FedSA-LoRA at $\tau = 10$ and $\tau = 20$ (by 3.19 and 0.96 points). On QNLI, two-factor sharing is the most accurate at every τ, at twice the uplink, and the one-sided rules lie within 0.7 points of each other for $\tau \geq 20$.

Table 7: Final accuracy versus the number of local steps per round τ (rounds $T = 3200/\tau$; $T = 3$ at $\tau = 1280$).

τ (rounds T)	10 (320)	20 (160)	40 (80)	80 (40)	160 (20)	320 (10)	640 (5)	1280 (3)
MNLI								
Local only	$89.64_{\pm 0.10}$	$90.12_{\pm 0.27}$	$90.00_{\pm 0.23}$	$90.55_{\pm 0.30}$	$90.07_{\pm 0.38}$	$90.12_{\pm 0.42}$	$90.03_{\pm 0.34}$	$88.94_{\pm 1.38}$
FedAvg-LoRA, personal head	**90.48**$_{\pm 0.17}$	**90.70**$_{\pm 0.31}$	**90.58**$_{\pm 0.25}$	**90.60**$_{\pm 0.08}$	$90.24_{\pm 0.48}$	$90.05_{\pm 0.35}$	$90.20_{\pm 0.14}$	$89.86_{\pm 0.31}$
FedSA-LoRA	$89.90_{\pm 0.05}$	$90.39_{\pm 0.11}$	$90.24_{\pm 0.13}$	$90.08_{\pm 0.24}$	**90.51**$_{\pm 0.42}$	$90.36_{\pm 0.27}$	**90.38**$_{\pm 0.30}$	**90.14**$_{\pm 0.20}$
PERSIA	$86.71_{\pm 1.05}$	$89.43_{\pm 0.55}$	$89.69_{\pm 0.66}$	$90.14_{\pm 0.33}$	$90.35_{\pm 0.32}$	**90.46**$_{\pm 0.08}$	$90.30_{\pm 0.29}$	$90.13_{\pm 0.30}$
QNLI								
Local only	$91.18_{\pm 0.54}$	$91.15_{\pm 0.07}$	$91.04_{\pm 1.19}$	$90.38_{\pm 0.79}$	$91.34_{\pm 0.13}$	$90.07_{\pm 1.79}$	$90.82_{\pm 1.56}$	$91.64_{\pm 0.40}$
FedAvg-LoRA, personal head	**93.37**$_{\pm 0.29}$	**93.45**$_{\pm 0.24}$	**93.82**$_{\pm 0.09}$	**93.37**$_{\pm 0.24}$	**93.67**$_{\pm 0.37}$	**94.00**$_{\pm 0.15}$	**93.18**$_{\pm 0.57}$	**93.05**$_{\pm 0.31}$
FedSA-LoRA	$92.01_{\pm 0.39}$	$91.95_{\pm 0.17}$	$91.72_{\pm 0.53}$	$91.51_{\pm 0.31}$	$91.88_{\pm 0.20}$	$92.11_{\pm 0.15}$	$92.18_{\pm 0.25}$	$92.12_{\pm 0.25}$
PERSIA	$90.74_{\pm 0.63}$	$92.11_{\pm 0.48}$	$92.39_{\pm 0.15}$	$92.15_{\pm 0.41}$	$92.09_{\pm 0.19}$	$91.84_{\pm 0.14}$	$92.00_{\pm 0.23}$	$92.14_{\pm 0.20}$

At $\tau = 640$ ($T = 5$), all five tasks are reported in Table 8: the federated rules lie within 0.24 points of each other on average.

Table 8: Final accuracy on all five tasks at $\tau = 640$, $T = 5$.

Method	MNLI	QNLI	QQP	SST-2	RTE	Avg.
Local only	$90.03_{\pm 0.34}$	$90.82_{\pm 1.56}$	$88.81_{\pm 0.09}$	$96.10_{\pm 0.35}$	$85.32_{\pm 0.62}$	91.44
FedAvg-LoRA, personal head	$90.20_{\pm 0.14}$	**93.18**$_{\pm 0.57}$	$88.84_{\pm 0.51}$	$96.22_{\pm 0.34}$	$85.42_{\pm 0.74}$	**92.11**
FedSA-LoRA	**90.38**$_{\pm 0.30}$	$92.18_{\pm 0.25}$	$89.16_{\pm 0.12}$	**96.58**$_{\pm 0.76}$	**86.03**$_{\pm 1.06}$	92.07
PERSIA	$90.30_{\pm 0.29}$	$92.00_{\pm 0.23}$	**89.22**$_{\pm 0.22}$	$95.95_{\pm 0.30}$	$85.39_{\pm 1.15}$	91.87

C.3 ACCURACY RIGHT AFTER AGGREGATION

Table 9 evaluates the models right after the first and the second aggregation, before any further local training; local only is the clients' model after the same local training. All rules start from the same seeds and data, so their models before the first aggregation agree up to GPU nondeterminism.

- **First aggregation.** PERSIA and PERSIA-R stay within 1.94 and 2.05 points of local training and exceed it at $\tau = 1280$, whereas factor averaging loses 3.09 to 10.25 points and two-factor averaging 12.96 to 26.12 points. Rebasing recovers part of the loss of factor averaging, and whitening adds a further 3.81 to 4.13 points on MNLI and 0.64 to 3.10 points on QNLI.
- **Second aggregation.** The other rules have recovered most of their loss. PERSIA or PERSIA-R is the most accurate federated rule on MNLI, and two-factor sharing, which uploads both factors, on QNLI at $\tau \geq 640$.

Table 9: Accuracy right after the first and the second aggregation, with τ local steps per round.

	MNLI			QNLI		
Method	$\tau = 320$	640	1280	320	640	1280
After the first aggregation (one local phase of τ steps)						
Local only (no aggregation)	**85.20**$_{\pm 0.80}$	**88.19**$_{\pm 0.35}$	88.63$_{\pm 1.42}$	86.98$_{\pm 2.31}$	**89.84**$_{\pm 1.43}$	90.87$_{\pm 0.60}$
FedAvg-LoRA, personal head	64.93$_{\pm 1.36}$	62.07$_{\pm 2.53}$	65.88$_{\pm 4.33}$	73.02$_{\pm 1.80}$	76.88$_{\pm 1.41}$	77.10$_{\pm 4.15}$
FedSA-LoRA	77.95$_{\pm 2.20}$	77.94$_{\pm 1.14}$	80.27$_{\pm 4.13}$	80.39$_{\pm 1.46}$	86.75$_{\pm 0.45}$	86.99$_{\pm 2.49}$
Factor avg. + rebasing	80.82$_{\pm 1.48}$	82.12$_{\pm 1.01}$	85.44$_{\pm 1.16}$	84.27$_{\pm 0.87}$	88.97$_{\pm 0.12}$	89.53$_{\pm 1.46}$
PERSIA-R‡	84.72$_{\pm 0.75}$	86.14$_{\pm 1.86}$	**89.34**$_{\pm 0.07}$	87.33$_{\pm 1.64}$	89.55$_{\pm 0.38}$	**91.34**$_{\pm 0.10}$
PERSIA	84.63$_{\pm 0.78}$	86.25$_{\pm 1.42}$	89.28$_{\pm 0.16}$	**87.37**$_{\pm 1.79}$	89.61$_{\pm 0.62}$	91.13$_{\pm 0.43}$
After the second aggregation (two local phases)						
Local only (no aggregation)	**88.14**$_{\pm 0.06}$	89.36$_{\pm 0.43}$	88.73$_{\pm 1.32}$	89.08$_{\pm 1.85}$	90.47$_{\pm 1.50}$	91.26$_{\pm 0.65}$
FedAvg-LoRA, personal head	86.16$_{\pm 0.84}$	87.19$_{\pm 0.75}$	88.63$_{\pm 0.12}$	90.73$_{\pm 0.17}$	**92.48**$_{\pm 0.10}$	**92.87**$_{\pm 0.17}$
FedSA-LoRA	86.76$_{\pm 0.55}$	88.01$_{\pm 0.70}$	88.78$_{\pm 0.21}$	90.39$_{\pm 0.22}$	91.41$_{\pm 0.36}$	91.51$_{\pm 0.28}$
Factor avg. + rebasing	87.29$_{\pm 0.35}$	88.78$_{\pm 0.16}$	89.68$_{\pm 0.16}$	91.05$_{\pm 0.39}$	91.29$_{\pm 0.32}$	91.81$_{\pm 0.54}$
PERSIA-R‡	87.97$_{\pm 0.04}$	89.27$_{\pm 0.24}$	**90.08**$_{\pm 0.29}$	**91.08**$_{\pm 0.48}$	90.68$_{\pm 0.18}$	91.77$_{\pm 0.28}$
PERSIA	88.08$_{\pm 0.28}$	**89.40**$_{\pm 0.33}$	89.77$_{\pm 0.23}$	90.77$_{\pm 0.38}$	91.39$_{\pm 0.19}$	92.10$_{\pm 0.21}$

C.4 ROBUSTNESS

Label skew, client count, and rank. These settings are reported in Table 3 of the main text. At $\alpha = 0.1$, the per-client majority floor is about 95% on MNLI and 97% on QNLI, so all methods lie in a narrow band. At $r = 4$, individual seeds of both one-sided methods collapse to the majority class, which explains their large spread.

Partial participation. In Table 10, k of the 10 clients report in each round; the others still rebase onto the broadcast right factor (Algorithm 1). With $k = 3$ on MNLI, PERSIA is 0.61 points more accurate than FedSA-LoRA with a third of its seed spread (0.64 versus 1.90).

Table 10: Partial participation: k of 10 clients per round, $\tau = 40$, $T = 80$.

	MNLI		QNLI		
Method	$k = 3$	$k = 5$	$k = 3$	$k = 5$	$k = 10$
FedAvg-LoRA, personal head	90.02$_{\pm 0.31}$	90.41$_{\pm 0.37}$	93.69$_{\pm 0.12}$	93.71$_{\pm 0.38}$	93.82$_{\pm 0.09}$
FedSA-LoRA	87.44$_{\pm 1.90}$	89.50$_{\pm 0.26}$	91.41$_{\pm 0.70}$	92.09$_{\pm 0.15}$	91.72$_{\pm 0.53}$
PERSIA	88.05$_{\pm 0.64}$	88.81$_{\pm 0.46}$	91.19$_{\pm 0.16}$	91.85$_{\pm 0.46}$	92.39$_{\pm 0.15}$

Compute heterogeneity. In Table 11, each QNLI client runs its own number of local steps per round ($T = 40$). In uni64 and uni88, every client takes 64 or 88 steps; in mild, severe, and extreme, the clients cycle through $\{20, 40, 80\}$, $\{10, 40, 160\}$, and $\{5, 20, 80, 320\}$ steps (means 44, 64, and 87.5). PERSIA matches or exceeds FedSA-LoRA with uniform or mildly heterogeneous budgets, but under severe and extreme heterogeneity it drops from 92.27 and 92.43 (uniform budgets with the same mean) to 88.29 and 88.48. The loss is thus caused by the spread of budgets rather than by the amount

of local work: the stacked objective weights each client by its update energy (Proposition 3.1), so clients that take many more steps dominate the shared subspace. Normalizing the stacked blocks (Remark F.6) is a natural remedy.

Table 11: Compute heterogeneity on QNLI: per-client local-step budgets, $T = 40$.

Method	uni64	uni88	mild	severe	extreme
FedAvg-LoRA, personal head	**93.43**$_{\pm 0.21}$	**94.08**$_{\pm 0.25}$	**93.61**$_{\pm 0.12}$	**93.71**$_{\pm 0.32}$	**93.70**$_{\pm 0.36}$
FedSA-LoRA	$92.10_{\pm 0.43}$	$91.99_{\pm 0.33}$	$91.79_{\pm 0.73}$	$91.33_{\pm 0.35}$	$91.62_{\pm 0.77}$
PERSIA	$92.27_{\pm 0.21}$	$92.43_{\pm 0.24}$	$91.93_{\pm 0.39}$	$88.29_{\pm 0.70}$	$88.48_{\pm 0.70}$

Heterogeneous tasks. In the six-task mixture of Section 6, 12 clients form pairs, and each pair holds one of QNLI, QQP, SST-2, RTE, MRPC, and CoLA, split IID within the pair ($\tau = 40$, $T = 80$). The score is the mean personalized accuracy over all clients: PERSIA reaches $85.79_{\pm 0.50}$, PERSIA-R $85.58_{\pm 0.58}$, and FedSA-LoRA $84.93_{\pm 0.16}$.

D COMMUNICATION AND COMPUTATION COST

D.1 COMMUNICATION

Table 12 reports the analytic LoRA payload per client and round in bf16 over all adapted layers of RoBERTa-Large, excluding the classification heads, which stay local, and protocol metadata. PERSIA uploads the right factor and an $r \times r$ Gram matrix per layer, i.e., essentially the uplink of FedSA-LoRA and half that of two-factor sharing.

Table 12: Per-client LoRA payload per round in the personalized protocol (bf16, MiB).

Method	Up / round	Down / round	Up, 80 rounds
Local only	0.00	0.00	0
FedSA-LoRA, factor avg. + rebase, FFA-LoRA	0.75	0.75	60
PERSIA, PERSIA-R	0.76	0.75	60
FedAvg-/FedEx-LoRA, FedProx (personal head)	1.50	1.50	120

D.2 SERVER COMPLEXITY

We count the per-round server cost for one adapted $d \times d$ matrix with N clients and rank $r \ll d$, assuming all client messages are held in memory; for L adapted matrices, all costs scale by L. A dense product of an $m \times n$ and an $n \times p$ matrix costs $2mnp$ FLOPs.

- **Full-rank averaging with SVD** reconstructs and averages the N products B_iA_i ($2Nd^2r$ FLOPs) and truncates the dense average by SVD ($O(d^2r)$), with $2Ndr + d^2$ floats of memory.
- **Factor-wise averaging** averages the factors in $O(Ndr)$ time and memory.
- **PERSIA** forms the Gram matrices and whitened factors ($4Ndr^2$), their square roots ($O(Nr^3)$), the truncated SVD of the $(Nr) \times d$ stack ($O(Ndr^2)$), and the rebased factors ($4Ndr^2$), i.e., $O(Ndr^2)$ time and $O(Ndr)$ memory. In personalized FL, whitening and rebasing run on the clients, and the server stores only the stack (Nrd floats).
- **Exact core SVD** (global FL only) computes the exact rank-r truncation of $\bar{W} = \frac{1}{N}\sum_i B_iA_i$ without forming it: thin QR factorizations of the stacked factors $[B_1|\cdots|B_N]$ and $[A_1;\ldots;A_N]^\top$ reduce the problem to the SVD of an $(Nr) \times (Nr)$ core, at $4d(Nr)^2 + O((Nr)^3)$ FLOPs and $O(Ndr + (Nr)^2)$ memory. It is exact and preferable when both factors are available and Nr is moderate, but its cost is quadratic in N.

Table 13 summarizes the scalings. Table 14 instantiates the leading terms for a 4096×4096 matrix with $r = 8$, using a randomized truncated SVD for PERSIA ($\approx 16Ndr^2$ FLOPs). Our experiments use an exact thin SVD of the stack, which costs $O(d(Nr)^2)$ and is negligible at $N = 10$ (Appendix D.3).

Table 13: Per-round server complexity for one adapted matrix, with all client messages held in memory.

Method	Time	Memory	Needs B_i at the server
Factor-wise averaging	$\mathcal{O}(Ndr)$	$\mathcal{O}(Ndr)$	in global FL only
PERSIA	$\mathcal{O}(Ndr^2)$	$\mathcal{O}(Ndr)$	no
Full-rank averaging + SVD	$\mathcal{O}(Nd^2r + d^2r)$	$\mathcal{O}(Ndr + d^2)$	yes
Exact core SVD	$\mathcal{O}(d(Nr)^2 + (Nr)^3)$	$\mathcal{O}(Ndr + (Nr)^2)$	yes

Table 14: Leading-order server FLOPs per adapted 4096×4096 matrix and round ($r = 8$). The core-SVD count includes only the two thin QR factorizations and therefore underestimates its cost.

Clients N	Core width Nr	Exact core SVD, $4d(Nr)^2$	PERSIA, $16Ndr^2$
10	80	0.10 GFLOPs	0.04 GFLOPs
100	800	10.5 GFLOPs	0.42 GFLOPs
500	4,000	262 GFLOPs	2.10 GFLOPs

D.3 WALL-CLOCK OVERHEAD

Tables 15 and 16 report per-round wall-clock times at 40 local steps, measured with an earlier implementation of the same aggregation rule; the server time covers the end-to-end aggregation. The server step of PERSIA costs less than 2% of a local round on RoBERTa-Large/GLUE and less than 0.2% on Qwen2.5-3B/GSM8K, and adds at most 0.03 s per round over factor-wise averaging.

Table 15: Per-round wall-clock time on RoBERTa-Large/GLUE at 40 local steps.

Method	Server (s)	Client (s)	Server / Client	vs. factor-wise
Global FL				
FFA-LoRA	0.022	3.16	0.70%	−0.010s
Factor-wise	0.032	3.28	0.98%	— (baseline)
FedProx	0.033	4.28	0.77%	+0.001s
FedMA	0.038	3.23	1.18%	+0.006s
PERSIA (w/o whitening)	0.061	3.26	1.87%	+0.029s
PERSIA	**0.062**	**3.25**	**1.91%**	**+0.030**s
Full-rank + SVD	0.076	3.29	2.31%	+0.044s
Personalized FL				
FedProx	0.023	4.14	0.56%	−0.002s
Factor-wise	0.025	3.25	0.77%	— (baseline)
FedMA	0.032	3.22	0.99%	+0.007s
PERSIA (w/o whitening)	0.052	3.31	1.57%	+0.027s
PERSIA	**0.054**	**3.34**	**1.62%**	**+0.030**s

Table 16: Per-round wall-clock time on Qwen2.5-3B/GSM8K at 40 local steps.

Method	Server (s)	Client (s)	Server / Client	vs. factor-wise
Global FL				
FFA-LoRA	0.018	14.81	0.12%	−0.005s
FedMA	0.019	14.79	0.13%	−0.004s
FedProx	0.022	15.67	0.14%	−0.001s
Full-rank + SVD	0.022	14.83	0.15%	−0.001s
Factor-wise	0.023	14.84	0.15%	— (baseline)
PERSIA (w/o whitening)	0.024	14.81	0.16%	+0.001s
PERSIA	**0.024**	**14.86**	**0.16%**	**+0.001**s
Personalized FL				
FedMA	0.020	14.79	0.14%	−0.002s
PERSIA (w/o whitening)	0.021	17.77	0.12%	−0.001s
FedProx	0.022	15.67	0.14%	+0.000s
Factor-wise	0.022	14.82	0.15%	— (baseline)
PERSIA	**0.023**	**15.23**	**0.13%**	**+0.001**s

E PERSIA IN GLOBAL FEDERATED LEARNING

In the global (two-factor) protocol, every client uploads both A_i and B_i, and the server broadcasts a single adapter. Exact rank-r aggregation is then available, by dense reconstruction or by the core SVD of Appendix D.2, and should be preferred whenever its cost is acceptable. The global variant of PERSIA (Algorithm 2, Figure 5) is a purely low-rank surrogate whose error is controlled by the stacked residual $\mathcal{R}$ and by the coherence of the clients' left subspaces. This appendix gives the aggregation rule (Appendix E.1), its error bounds (Appendix E.2), a remark on convergence (Appendix E.3), and experiments (Appendix E.4).

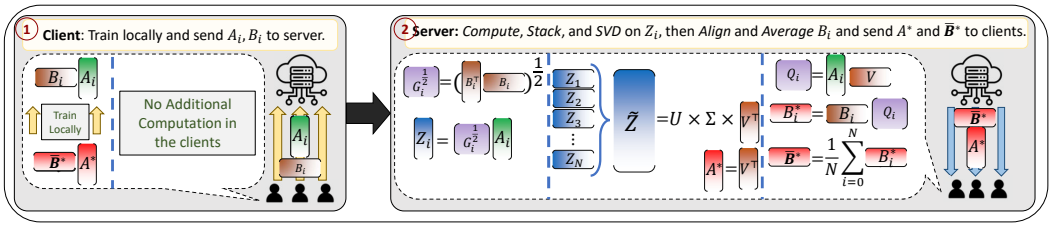


Figure 5: PERSIA in global FL: Z_i, $A^\star$, and $B^\star$ are computed at the server.

E.1 AGGREGATION RULE

The server computes Z_i from (A_i, B_i), forms $\tilde{Z}$ as in equation 5, and computes V_r as in equation 6. It then rebases each left factor onto the new basis with $Q_i := A_i V_r \in \mathbb{R}^{r\times r}$, $B_i' = B_i Q_i$, averages the B_i' into $B^\star$, and broadcasts $(B^\star, V_r^\top)$. The polar factor O_i of B_i is never formed, since $B_i' = O_i(Z_i V_r) = O_i G_i A_i V_r = B_i A_i V_r = B_i Q_i$. The aggregated update is a right projection of the full-rank average,

$$B^\star A^\star = \Big(\frac{1}{N}\sum_{i=1}^{N} B_i'\Big)V_r^\top = \Big(\frac{1}{N}\sum_{i=1}^{N} B_i A_i\Big)P_r, \qquad P_r := V_r V_r^\top, \tag{10}$$

where P_r is an orthogonal projection, $P_r^2 = V_r(V_r^\top V_r)V_r^\top = P_r$.

E.2 AGGREGATION ERROR BOUNDS AND OPTIMALITY REGIMES

This subsection states what PERSIA is optimal for in global FL. Three objectives must be distinguished: (i) the stacked whitened right-space objective that PERSIA solves, (ii) the personalized client-wise projection objective, and (iii) the dense global-FL oracle. The first two coincide (Proposition 3.1); the third coincides with them in several regimes, but not in general. Proofs are given in Appendix F; the regime labels in Table 17 are heuristics rather than theorems.

Throughout this section we consider one adapted matrix; the multilayer case is obtained by summing the corresponding residuals over layers. For every B_i, let

$$\Delta W_i = B_i A_i = O_i Z_i, \qquad O_i^\top O_i = I_r, \qquad Z_i := G_i A_i,$$

be the Gram-whitened representation of client i, where O_i is a polar factor of B_i; it exists for every B_i (Lemma F.7). Define

$$\widetilde{Z} := \begin{bmatrix} Z_1 \\ \vdots \\ Z_N \end{bmatrix}, \qquad \mathcal{P}_Z := \{P = V_r V_r^\top : \; V_r^\top V_r = I_r\}.$$

PERSIA computes

$$P_Z := V_r V_r^\top,$$

where V_r contains the top-r right singular vectors of $\widetilde{Z}$. Equivalently,

$$P_Z \in \arg\min_{P\in\mathcal{P}_Z} \|\widetilde{Z}(I-P)\|_F^2.$$

We denote the optimal stacked residual by

$$\mathcal{R}_Z := \|\widetilde{Z}(I - P_Z)\|_F^2 = \sum_{j>r} \sigma_j^2(\widetilde{Z}).$$

Definition E.1 (Aggregation errors). *Let $\overline{W} := \frac{1}{N}\sum_{i=1}^N \Delta W_i$, denote the hypothetical dense full-rank average. For a global aggregated update $B'A'$, define aggregation error as:*

$$E_{\mathrm{FL}}(A', B') := \|\overline{W} - B'A'\|_F.$$

The personalized error E_{PFL} is defined in equation 2.

Proposition E.2 (PERSIA error in FL). *Given $\tilde{Z}$ and V_r as in equation 5 and equation 6, and having $A^\star = V_r^\top$, and $B^\star = \frac{1}{N}\sum_{i=1}^N B_i(A_i V_r)$. Then we have*

$$E_{\mathbf{FL}}(A^\star, B^\star)^2 \le \frac{1}{N}\,\|\tilde{Z} - \tilde{Z}V_r V_r^\top\|_F^2 = \frac{\mathcal{R}}{N},$$

where $\mathcal{R} := \sum_{j>r}\sigma_j^2(\tilde{Z}) = \min_{V_r^\top V_r = I_r}\|\tilde{Z} - \tilde{Z}V_r V_r^\top\|_F^2$, denoting the stacked residual tail of $\tilde{Z}$.

Proposition E.3 (Global-FL error bound). *Let the global-FL PERSIA output be*

$$A^\star = V_r^\top, \qquad B^\star = \frac{1}{N}\sum_{i=1}^N B_i(A_i V_r).$$

Define the cross-client coherence

$$\mu := \max_{i\neq j}\|O_i^\top O_j\|_2,$$

and the block Gram matrix $K_O \in \mathbb{R}^{Nr\times Nr}$ with (i,j)-block $O_i^\top O_j$ (formally introduced in Proposition E.4 below). Then

$$\lambda_{\max}(K_O) \le 1 + (N-1)\mu,$$

and therefore

$$E_{\mathrm{FL}}(A^\star, B^\star)^2 \le \frac{1 + (N-1)\mu}{N^2}\mathcal{R}_Z.$$

In particular, since $\mu \le 1$, this recovers the coarse bound

$$E_{\mathrm{FL}}(A^\star, B^\star)^2 \le \frac{\mathcal{R}_Z}{N}.$$

Proposition E.4 (Global FL as a metric-weighted right-space problem). *Define*

$$\overline{O} := \frac{1}{N}[O_1|\cdots|O_N], \qquad K_O := (N\overline{O})^\top(N\overline{O}),$$

so that the (i,j)-block of K_O is $O_i^\top O_j$. Then

$$\overline{W} = \overline{O}\widetilde{Z},$$

and for any $P \in \mathcal{P}_Z$,

$$\|\overline{W}(I-P)\|_F^2 = \|\overline{O}\widetilde{Z}(I-P)\|_F^2 = \frac{1}{N^2}\,\mathrm{tr}\Big((\widetilde{Z}(I-P))^\top K_O \widetilde{Z}(I-P)\Big).$$

Thus the dense global-FL oracle uses the metric K_O, whereas PERSIA uses the Euclidean metric I on the stacked whitened factors.

Definition E.5 (PERSIA and dense-oracle global-FL residuals). *Let V_r^Z denote the top-r right singular vectors of $\widetilde{Z}$, and let V_r^W denote the top-r right singular vectors of $\overline{W}$. Define the squared PERSIA global-FL residual*

$$E_{\mathrm{PER}} := \|\overline{W}(I - V_r^Z(V_r^Z)^\top)\|_F^2,$$

and the squared dense-oracle residual

$$E_{\mathrm{ora}} := \min_{P\in\mathcal{P}_Z}\|\overline{W}(I-P)\|_F^2 = \sum_{j>r}\sigma_j^2(\overline{W}).$$

Proposition E.6 (Surrogate gap)**.** *With K_O as in Proposition E.4,*

$$0 \;\le\; E_{\mathrm{PER}} - E_{\mathrm{ora}} \;\le\; \frac{2r}{N^2}\,\big\|\widetilde{Z}^\top (K_O - I)\widetilde{Z}\big\|_2.$$

In particular, the gap is zero whenever the whitened left subspaces are mutually orthogonal ($K_O = I$).

Corollary E.7 (Exact and near-exact regimes)**.** *PERSIA is exact or near-exact in the following regimes.*

(i) Personalized FL. *PERSIA exactly minimizes the PFL client-wise projection objective by Proposition 3.1.*

(ii) Orthogonal whitened left subspaces. *If $O_i^\top O_j = 0$ for all $i \neq j$, then $K_O = I$, and*

$$\|\overline{W}(I-P)\|_F^2 = \frac{1}{N^2}\|\widetilde{Z}(I-P)\|_F^2 \quad \forall P \in \mathcal{P}_Z.$$

Hence PERSIA coincides with the dense global-FL rank-r right-projection oracle.

(iii) Shared functionality under $GL(r)$ symmetry. *If all clients represent the same dense update under invertible basis changes,*

$$\Delta W_i = BA = (BT_i)(T_i^{-1}A), \qquad T_i \in GL(r),$$

then PERSIA exactly recovers the ideal dense average:

$$E_{\mathrm{FL}}(A^\star, B^\star) = 0, \qquad E_{\mathrm{PFL}}(A^\star, B_i^\star) = 0.$$

(iv) Small spectral distortion. *Let*

$$\rho := \|K_O - I\|_2.$$

If $\rho < 1$, then for all $P \in \mathcal{P}_Z$,

$$\frac{1-\rho}{N^2}\|\widetilde{Z}(I-P)\|_F^2 \le \|\overline{W}(I-P)\|_F^2 \le \frac{1+\rho}{N^2}\|\widetilde{Z}(I-P)\|_F^2.$$

Therefore, if $E_{\mathrm{ora}} > 0$, then PERSIA satisfies the approximation ratio

$$\frac{E_{\mathrm{FL}}(A^\star, B^\star)^2}{E_{\mathrm{ora}}} \le \frac{1+\rho}{1-\rho}.$$

(v) Principal-angle and coherence-controlled oracle gap. *Let*

$$\delta_Z := \sigma_r^2(\widetilde{Z}) - \sigma_{r+1}^2(\widetilde{Z})$$

be the rank-r spectral gap of $\widetilde{Z}^\top\widetilde{Z}$. Then

$$E_{\mathrm{PER}} - E_{\mathrm{ora}} \le \sigma_1^2(\overline{W})\|\sin\Theta(V_r^W, V_r^Z)\|_F^2.$$

Moreover, by the Davis–Kahan theorem (Davis & Kahan, 1970), if

$$2(N-1)\mu\,\sigma_1^2(\widetilde{Z}) < \delta_Z,$$

then

$$E_{\mathrm{PER}} - E_{\mathrm{ora}} \le 4r\,\sigma_1^2(\overline{W})\left(\frac{(N-1)\mu\,\sigma_1^2(\widetilde{Z})}{\delta_Z}\right)^2.$$

Thus the gap is quadratically small in μ and decreases as the spectral gap δ_Z increases.

Low-rank regime. For generic independent r-dimensional subspaces of $\mathbb{R}^d$, $\mu = O(\sqrt{r/d})$, so the coherence factor of the oracle gap in Corollary E.7(v) scales as $\mu^2 = O(r/d)$; the gap also carries the rank, client-count, and spectral factors shown there, so this is a scaling guide rather than a dimension-only guarantee. The maximum over all $N(N-1)$ pairs adds a $\sqrt{\log N/d}$ term, and adapters trained from a common initialization can be far more coherent than random subspaces, so μ should be measured rather than assumed for trained adapters; Proposition E.6 bounds the gap by the directly measurable quantity $\|\widetilde{Z}^\top(K_O - I)\widetilde{Z}\|_2$.

Interpretation. PERSIA is always exactly optimal for the stacked whitened objective, which equals the personalized objective. In global FL, the dense oracle uses the metric K_O, which contains all pairwise interactions between the whitened left subspaces; PERSIA avoids estimating it and solves the Euclidean surrogate instead. The resulting gap is controlled by the stacked residual $\mathcal{R}_Z$ and by the coherence, or spectral distortion, of the whitened left subspaces. Table 17 summarizes the regimes.

Table 17: Regimes that control the relation of PERSIA to the dense oracle in global FL.

Regime	Coherence μ	PERSIA versus full-rank average-then-SVD
Personalized FL	not applicable	Exact client-wise optimum (full rank unavailable)
IID data	$\mu \to 1$	Loosest bound; exact under shared-function symmetry
Mild non-IID	moderate	Intermediate; potentially the hardest regime
Strong non-IID or multi-task	$\mu \to 0$	Near-optimal; exact in the orthogonal limit
Exact $GL(r)$ symmetry	not applicable	Zero error
Low rank, random subspaces	$\mathcal{O}(\sqrt{r/d})$	Coherence factor scales as $O(r/d)$

E.3 Convergence in global FL

The master inequality and the jump bound of Appendix G (Lemmas G.4 and G.5) hold for any aggregation rule and hence also for the global variant, where every client receives $\Theta^t_{\mathrm{avg}} P^t_r$ with $\Theta^t_{\mathrm{avg}} := \frac{1}{N}\sum_i \Delta W_i^{t,\tau}$. In global FL, however, the aggregation displacement of client i also contains its drift $\Theta^t_{\mathrm{avg}} - \Delta W_i^{t,\tau}$ from the average, which is of first order in $\eta\tau$ for every rule that returns a single model. Bounding this drift with the same worst-case constants yields a floor no smaller than the trivial bound $\mathcal{G}_i \le (\rho_A^2 + \rho_B^2) G_\nabla^2$ on $\mathcal{X}$, so we do not state a global-FL rate; a meaningful guarantee would measure stationarity of the average objective at the broadcast point, which we leave to future work. The deterministic results above (Propositions E.2, E.3 and E.6) are unaffected.

E.4 Experiments

Setup. On GLUE, we use the partitions, schedules, hyperparameters, and client-holdout evaluation of Section 5, but the server aggregates both adapter factors and the classification head, so nothing stays on the clients; "global" refers to what is aggregated, not to the evaluation data. We compare PERSIA (Algorithm 2) with FedAvg-LoRA (McMahan et al., 2017), FedEx-LoRA (Singhal et al., 2025), FFA-LoRA (Sun et al., 2024), FLoRA (Wang et al., 2024), FLoRIST (Ramesh & Dass, 2026), LoRA-FAIR (Bian et al., 2025), and dense average-then-SVD (Stoica et al., 2025). On GSM8K (Cobbe et al., 2021), we fine-tune Qwen2.5-3B-Instruct (Qwen Team, 2024) with 5 IID clients, evaluate generation accuracy on 128 held-out problems, and keep $\tau T \approx 300$ local steps.

Standard schedule. At $\tau = 40$ (Table 18), PERSIA attains the highest accuracy on MNLI and QNLI and is within 0.2 points of the best method on QQP and SST-2. On RTE, all methods that train both factors reach only 68–71%, whereas FFA-LoRA reaches 85.3%. Removing error feedback leaves PERSIA essentially unchanged ($86.26_{\pm 0.58}$ on MNLI and $93.14_{\pm 0.30}$ on QNLI).

Table 18: Global protocol on GLUE with RoBERTa-Large, $\tau = 40$, $T = 80$.

Method	MNLI	QNLI	QQP	SST-2	RTE
FedAvg-LoRA (McMahan et al., 2017)	$85.32_{\pm 0.66}$	$93.02_{\pm 0.41}$	$86.54_{\pm 0.04}$	$95.43_{\pm 0.45}$	$67.78_{\pm 0.26}$
FedEx-LoRA (Singhal et al., 2025)	$85.32_{\pm 0.48}$	$92.88_{\pm 0.34}$	$86.60_{\pm 0.17}$	$95.23_{\pm 0.20}$	$67.87_{\pm 2.04}$
FFA-LoRA (Sun et al., 2024)	$84.86_{\pm 0.90}$	$88.64_{\pm 0.44}$	$85.21_{\pm 0.29}$	$\mathbf{95.54_{\pm 0.17}}$	$\mathbf{85.26_{\pm 0.53}}$
FLoRA (Wang et al., 2024)	$85.57_{\pm 0.35}$	$92.99_{\pm 0.07}$	$\mathbf{86.67_{\pm 0.14}}$	$95.31_{\pm 0.16}$	$67.74_{\pm 1.80}$
Average-then-SVD (Stoica et al., 2025)	$85.78_{\pm 0.31}$	$92.90_{\pm 0.24}$	$86.53_{\pm 0.16}$	$95.39_{\pm 0.06}$	$67.73_{\pm 1.35}$
FLoRIST (Ramesh & Dass, 2026)	$85.30_{\pm 0.39}$	$92.56_{\pm 0.39}$	$86.39_{\pm 0.18}$	$95.10_{\pm 0.13}$	$68.98_{\pm 0.93}$
LoRA-FAIR (Bian et al., 2025)	$85.49_{\pm 0.60}$	$93.07_{\pm 0.32}$	$86.46_{\pm 0.15}$	$95.52_{\pm 0.33}$	$70.61_{\pm 2.10}$
PERSIA (Algorithm 2)	$\mathbf{86.30_{\pm 0.68}}$	$\mathbf{93.25_{\pm 0.17}}$	$86.58_{\pm 0.46}$	$95.35_{\pm 0.16}$	$70.22_{\pm 2.23}$

Long local horizon. At $\tau = 640$ (Table 19), PERSIA is the most accurate of the six methods on MNLI and QNLI, and FFA-LoRA collapses. At $\tau = 320$, PERSIA is ahead of FedAvg-LoRA and

average-then-SVD on MNLI ($86.86_{\pm 0.09}$ versus $86.16_{\pm 0.33}$ and $86.19_{\pm 0.11}$), and FedAvg-LoRA is ahead on QNLI ($93.45_{\pm 0.16}$ versus $93.10_{\pm 0.33}$). With strong label skew ($\alpha = 0.1$, $\tau = 640$), FedAvg-LoRA is more accurate than PERSIA ($82.74_{\pm 0.60}$ versus $82.49_{\pm 0.39}$ on MNLI and $85.57_{\pm 1.07}$ versus $84.48_{\pm 1.02}$ on QNLI).

Table 19: Global protocol at a long local horizon, $\tau = 640$, $T = 5$.

Method	MNLI	QNLI
FedAvg-LoRA	$86.74_{\pm 0.38}$	$92.69_{\pm 0.33}$
FedEx-LoRA	$86.56_{\pm 0.26}$	$92.63_{\pm 0.13}$
FFA-LoRA	$45.75_{\pm 0.13}$	$62.43_{\pm 0.30}$
FLoRA	$86.53_{\pm 0.05}$	$92.87_{\pm 0.14}$
Average-then-SVD	$86.24_{\pm 0.40}$	$93.14_{\pm 0.32}$
PERSIA	$\mathbf{86.91_{\pm 0.36}}$	$\mathbf{93.19_{\pm 0.26}}$

GSM8K. On GSM8K (Table 20), PERSIA improves over FedAvg-LoRA, FedEx-LoRA, LoRA-FAIR, and FFA-LoRA at every horizon but trails dense average-then-SVD and FLoRA at every horizon and FLoRIST at $\tau \leq 75$; these methods aggregate exactly or through the stacked factors. We therefore view the global variant as a competitive low-rank surrogate rather than a replacement for exact aggregation when the latter is affordable.

Table 20: GSM8K with Qwen2.5-3B-Instruct, 5 IID clients, global protocol ($\tau T \approx 300$).

Method	$\tau = 10$	$\tau = 30$	$\tau = 75$	$\tau = 150$
Average-then-SVD	$\mathbf{49.48_{\pm 1.33}}$	$\mathbf{50.26_{\pm 1.84}}$	$\mathbf{49.48_{\pm 0.37}}$	$\mathbf{45.83_{\pm 1.61}}$
FLoRA	$47.92_{\pm 0.37}$	$49.22_{\pm 0.64}$	$48.96_{\pm 1.33}$	$45.31_{\pm 1.10}$
FLoRIST	$48.18_{\pm 0.97}$	$\mathbf{50.26_{\pm 3.15}}$	$47.92_{\pm 0.37}$	$44.01_{\pm 0.97}$
PERSIA	$46.09_{\pm 2.21}$	$45.83_{\pm 1.47}$	$45.83_{\pm 3.15}$	$45.05_{\pm 1.95}$
LoRA-FAIR	$42.45_{\pm 1.33}$	$42.45_{\pm 0.97}$	$44.53_{\pm 1.28}$	$43.75_{\pm 1.28}$
FedAvg-LoRA	$42.97_{\pm 1.10}$	$42.45_{\pm 1.33}$	$44.27_{\pm 1.47}$	$42.71_{\pm 0.37}$
FedEx-LoRA	$41.93_{\pm 1.84}$	$42.45_{\pm 1.61}$	$44.01_{\pm 1.33}$	$42.97_{\pm 1.10}$
FFA-LoRA	$34.90_{\pm 1.95}$	$34.38_{\pm 1.91}$	$34.90_{\pm 0.97}$	$34.64_{\pm 0.97}$

F PROOFS OF PROPOSITIONS

This appendix provides complete, step-by-step proofs and justifications for all mathematical claims in Section 3 whose proofs are omitted from the main text, including Proposition 3.1 and Corollary 3.2, the identities Equations (3), (4), (6) and (7), and the global-FL statements Propositions E.2 and E.6 of Appendix E.

F.1 STANDING NOTATION AND ASSUMPTIONS

Throughout this appendix, we use the same notation as in Section 3. For a fixed LoRA-injected linear layer, each client $i \in \{1, \dots, N\}$ produces a rank-r update

$$\Delta W_i := B_i A_i, \qquad B_i \in \mathbb{R}^{d\times r},\ A_i \in \mathbb{R}^{r\times k}.$$

No rank assumption. For every B_i, the matrix $B_i^\top B_i$ is symmetric positive semidefinite and therefore has a unique symmetric positive semidefinite square root

$$G_i := (B_i^\top B_i)^{1/2}, \qquad G_i^2 = B_i^\top B_i, \qquad \|G_i\|_2 = \|B_i\|_2.$$

All personalized results below use only G_i and hold for rank-deficient B_i. An orthonormal polar factor O_i with $B_i = O_i G_i$ exists for every B_i (Lemma F.7) and equals $B_i G_i^{-1}$ when B_i has full column rank; it is used only in the geometric interpretation and in the global-FL refinements of Appendix E.2.

Stabilized square root. In practice we may compute $G_{i,\varepsilon} := (B_i^\top B_i + \varepsilon I_r)^{1/2}$ for $\varepsilon > 0$.

Justification. For any $\varepsilon > 0$ and any nonzero $x \in \mathbb{R}^r$,

$$x^\top (B_i^\top B_i + \varepsilon I_r)x = x^\top (B_i^\top B_i)x + \varepsilon x^\top x = \|B_i x\|_2^2 + \varepsilon \|x\|_2^2 > 0.$$

Therefore $B_i^\top B_i + \varepsilon I_r$ is symmetric positive definite and invertible. Hence $(B_i^\top B_i + \varepsilon I_r)^{1/2}$ exists uniquely as a symmetric positive definite matrix, and its inverse exists as well. This guarantees numerical stability for the same algebraic constructions. Since $G_{i,\varepsilon}^2 = G_i^2 + \varepsilon I_r$, the stabilized factor satisfies $Z_{i,\varepsilon}^\top Z_{i,\varepsilon} = \Delta W_i^\top \Delta W_i + \varepsilon A_i^\top A_i$. Hence, if P_ε is the top-r right singular projector of the stabilized stack and $P^\star$ that of $\tilde{Z}$, then $\frac{1}{N}\sum_i \|\Delta W_i(I - P_\varepsilon)\|_F^2 \le \frac{1}{N}\sum_i \|Z_{i,\varepsilon}(I - P_\varepsilon)\|_F^2 \le \frac{1}{N}\sum_i \|Z_{i,\varepsilon}(I - P^\star)\|_F^2 \le \frac{\mathcal{R}}{N} + \frac{\varepsilon}{N}\sum_i \|A_i\|_F^2$, so the ridge costs at most $\frac{\varepsilon}{N}\sum_i \|A_i\|_F^2$ in the objective of Proposition 3.1; our analysis uses $\varepsilon = 0$. □

F.2 AUXILIARY LEMMAS USED BY PERSIA

Lemma F.1 (PERSIA Gram-whitening identities)**.** *Fix a client i and an arbitrary B_i. Define G_i as above and*

$$Z_i := G_i A_i.$$

Then $Z_i^\top Z_i = \Delta W_i^\top \Delta W_i$, and for every matrix M with k rows,

$$\|\Delta W_i M\|_F^2 = \|Z_i M\|_F^2. \tag{11}$$

In particular, the identities in the main text hold:

$$\|\Delta W_i\|_F^2 = \operatorname{tr}(A_i^\top B_i^\top B_i A_i) = \|G_i A_i\|_F^2 = \|Z_i\|_F^2, \tag{12}$$

$$\|\Delta W_i(I - P)\|_F^2 = \|Z_i(I - P)\|_F^2 \quad \textit{for any } P = P^\top = P^2 \in \mathbb{R}^{k\times k}, \tag{13}$$

which are exactly Equations (3) *and* (4)*.*

Proof. Recall $\Delta W_i = B_i A_i$ and $G_i^2 = B_i^\top B_i$. Since G_i is symmetric,

$$Z_i^\top Z_i = A_i^\top G_i^\top G_i A_i = A_i^\top G_i^2 A_i = A_i^\top B_i^\top B_i A_i = \Delta W_i^\top \Delta W_i.$$

For any M with k rows, using the Frobenius norm definition,

$$\|\Delta W_i M\|_F^2 = \operatorname{tr}\big(M^\top \Delta W_i^\top \Delta W_i M\big) = \operatorname{tr}\big(M^\top Z_i^\top Z_i M\big) = \|Z_i M\|_F^2,$$

which is equation 11. Taking $M = I$ gives equation 12, since $\operatorname{tr}(A_i^\top B_i^\top B_i A_i) = \operatorname{tr}(A_i^\top G_i^2 A_i) = \|G_i A_i\|_F^2$, and taking $M = I - P$ gives equation 13. No invertibility of G_i is used. □

Lemma F.2 (Stacking identity). *Let $Z_1, \dots, Z_N$ be matrices with the same number of columns, and define the vertical stack*

$$\tilde{Z} := \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_N \end{bmatrix}.$$

Then for any matrix M whose number of rows matches the common number of columns of the Z_i,

$$\|\tilde{Z}M\|_F^2 = \sum_{i=1}^{N} \|Z_i M\|_F^2.$$

In particular, if $M = I - P$ then

$$\|\tilde{Z}(I-P)\|_F^2 = \sum_{i=1}^{N} \|Z_i (I-P)\|_F^2.$$

Proof. By the definition of block matrix multiplication,

$$\tilde{Z}M = \begin{bmatrix} Z_1 M \\ Z_2 M \\ \vdots \\ Z_N M \end{bmatrix}.$$

By the definition of Frobenius norm squared as the sum of squares of all entries, the squared Frobenius norm of a vertical stack equals the sum of the squared Frobenius norms of the blocks:

$$\left\| \begin{bmatrix} Z_1 M \\ \vdots \\ Z_N M \end{bmatrix} \right\|_F^2 = \sum_{i=1}^{N} \|Z_i M\|_F^2.$$

□

Lemma F.3 (Least-squares rebasing). *Let $W \in \mathbb{R}^{d\times k}$, let $A' \in \mathbb{R}^{r\times k}$ be arbitrary, and let $P_{A'}$ be the orthogonal projector onto the row space of A', so that* $\operatorname{rank}(P_{A'}) \le r$. *Then*

$$\min_{C\in\mathbb{R}^{d\times r}} \|W - CA'\|_F^2 = \|W(I - P_{A'})\|_F^2,$$

attained at $C = W(A')^+$. If A' has orthonormal rows, $P_{A'} = A'^\top A'$ and the minimizer is $C = WA'^\top$.

Proof. Every row of CA' lies in the row space of A', so $CA'P_{A'} = CA'$. Hence

$$W - CA' = W(I - P_{A'}) + \big(WP_{A'} - CA'\big),$$

where the rows of the first term are orthogonal to the row space of A' and those of the second lie in it. The cross term vanishes, $\operatorname{tr}\big((I-P_{A'})W^\top(WP_{A'}-CA')\big) = \operatorname{tr}\big(W^\top(WP_{A'}-CA')(I-P_{A'})\big) = 0$, so

$$\|W - CA'\|_F^2 = \|W(I-P_{A'})\|_F^2 + \|WP_{A'} - CA'\|_F^2 \ge \|W(I-P_{A'})\|_F^2.$$

Equality holds for $C = W(A')^+$, because $(A')^+A' = P_{A'}$. If $A'A'^\top = I_r$, then $(A')^+ = A'^\top$ and $P_{A'} = A'^\top A'$. □

F.3 Proof of Proposition E.2

Proposition F.4 (PERSIA error in FL). *Given $\tilde{Z}$ and V_r as in equation 5 and equation 6, and having $A^\star = V_r^\top$, and $B^\star = \frac{1}{N}\sum_{i=1}^{N} B_i(A_i V_r)$. Then*

$$E_{\mathbf{FL}}(A^\star, B^\star)^2 \le \frac{1}{N}\,\|\tilde{Z} - \tilde{Z}V_r V_r^\top\|_F^2 = \frac{\mathcal{R}}{N},$$

where $\mathcal{R} := \sum_{j>r}\sigma_j^2(\tilde{Z}) = \min_{V_r^\top V_r = I_r} \|\tilde{Z} - \tilde{Z}V_r V_r^\top\|_F^2$.

Proof. We prove the inequality and the identification of the residual with the singular-value tail. In Proposition E.2, PERSIA sets

$$A^\star := V_r^\top, \qquad B^\star := \frac{1}{N}\sum_{i=1}^{N} B_i(A_i V_r).$$

Therefore

$$B^\star A^\star = \left(\frac{1}{N}\sum_{i=1}^{N} B_i(A_i V_r)\right) V_r^\top = \frac{1}{N}\sum_{i=1}^{N} B_i(A_i V_r) V_r^\top$$
$$= \frac{1}{N}\sum_{i=1}^{N} B_i A_i (V_r V_r^\top) = \left(\frac{1}{N}\sum_{i=1}^{N} \Delta W_i\right)(V_r V_r^\top).$$

By definition, $\overline{W} := \frac{1}{N}\sum_{i=1}^{N} \Delta W_i$, so

$$B^\star A^\star = \overline{W}(V_r V_r^\top).$$

By the definition of $E_{\mathbf{FL}}$,

$$E_{\mathbf{FL}}(A^\star, B^\star) = \|\overline{W} - B^\star A^\star\|_F = \|\overline{W} - \overline{W}(V_r V_r^\top)\|_F = \|\overline{W}(I - V_r V_r^\top)\|_F.$$

Thus

$$E_{\mathbf{FL}}(A^\star, B^\star)^2 = \left\|\overline{W}(I - V_r V_r^\top)\right\|_F^2.$$

Let $X_i := \Delta W_i(I - V_r V_r^\top)$. Then

$$\overline{W}(I - V_r V_r^\top) = \left(\frac{1}{N}\sum_{i=1}^{N} \Delta W_i\right)(I - V_r V_r^\top) = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

By the triangle inequality and Cauchy–Schwarz, we have

$$\left\|\frac{1}{N}\sum_{i=1}^{N} X_i\right\|_F \le \frac{1}{N}\sum_{i=1}^{N} \|X_i\|_F \le \left(\frac{1}{N}\sum_{i=1}^{N} \|X_i\|_F^2\right)^{1/2}.$$

Squaring both sides yields the desired bound:

$$\left\|\frac{1}{N}\sum_{i=1}^{N} X_i\right\|_F^2 \le \frac{1}{N}\sum_{i=1}^{N} \|X_i\|_F^2.$$

Recall $X_i = \Delta W_i(I - V_r V_r^\top)$. By Lemma F.1 with $P = V_r V_r^\top$,

$$\|\Delta W_i(I - V_r V_r^\top)\|_F^2 = \|Z_i(I - V_r V_r^\top)\|_F^2.$$

Therefore

$$\frac{1}{N}\sum_{i=1}^{N} \|X_i\|_F^2 = \frac{1}{N}\sum_{i=1}^{N} \|Z_i(I - V_r V_r^\top)\|_F^2.$$

By Lemma F.2 with $M = I - V_r V_r^\top$,

$$\sum_{i=1}^{N} \|Z_i(I - V_r V_r^\top)\|_F^2 = \|\tilde{Z}(I - V_r V_r^\top)\|_F^2 = \|\tilde{Z} - \tilde{Z} V_r V_r^\top\|_F^2.$$

Hence

$$E_{\mathbf{FL}}(A^\star, B^\star)^2 = \left\|\frac{1}{N}\sum_{i=1}^{N} X_i\right\|_F^2 \le \frac{1}{N}\|\tilde{Z} - \tilde{Z} V_r V_r^\top\|_F^2.$$

By the optimality of the SVD solution, since V_r is the optimizer of Equation (6),

$$\|\tilde{Z} - \tilde{Z}V_rV_r^\top\|_F^2 = \min_{V_r^\top V_r = I_r} \|\tilde{Z} - \tilde{Z}V_rV_r^\top\|_F^2 = \sum_{j>r} \sigma_j^2(\tilde{Z}).$$

Defining $\mathcal{R} := \sum_{j>r} \sigma_j^2(\tilde{Z})$ yields exactly the claimed bound:

$$E_{\mathbf{FL}}(A^\star, B^\star)^2 \le \frac{1}{N}\|\tilde{Z} - \tilde{Z}V_rV_r^\top\|_F^2 = \frac{\mathcal{R}}{N}.$$

□

F.4 Proof of Proposition 3.1

Proposition F.5 (Optimal shared right subspace in PFL)**.** *Given $\tilde{Z}$ and V_r as in equation 5 and equation 6 with arbitrary B_i and A_i, for every $A' \in \mathbb{R}^{r\times k}$ and all $\{B_i'\}_{i=1}^N$,*

$$E_{\mathbf{PFL}}(A', \{B_i'\}_{i=1}^N)^2 \;\ge\; \frac{\mathcal{R}}{N} \;=\; E_{\mathbf{PFL}}(A^\star, \{B_i^\star\}_{i=1}^N)^2,$$

where $A^\star = V_r^\top$, $B_i^\star = B_i(A_iV_r)$, and $\mathcal{R} := \sum_{j>r}\sigma_j^2(\tilde{Z}) = \min_{V_r^\top V_r = I_r}\|\tilde{Z} - \tilde{Z}V_rV_r^\top\|_F^2$. Moreover, $\mathcal{R}$ and the selected subspace depend only on $\{\Delta W_i\}$.

Proof. The proof has three steps: an attainment identity for PERSIA, a matching lower bound for every shared right factor, and the invariance statement.

Step 1: PERSIA attains $\mathcal{R}/N$. In Proposition 3.1, the notation $B_i^\star = B_i(A_iV_r)$ is client-dependent; in this proof we write it as

$$B_i^\star := B_i(A_iV_r), \qquad A^\star := V_r^\top.$$

By definition, as in the main text,

$$B_i^\star A^\star = \big(B_i(A_iV_r)\big)V_r^\top = \Delta W_i(V_rV_r^\top).$$

Using the definition of $E_{\mathbf{PFL}}$ with the client-specific choice $B_i' = B_i^\star$ and shared $A' = A^\star$,

$$\begin{aligned} E_{\mathbf{PFL}}(A^\star, \{B_i^\star\}_{i=1}^N)^2 &= \frac{1}{N}\sum_{i=1}^N \|\Delta W_i - B_i^\star A^\star\|_F^2 \\ &= \frac{1}{N}\sum_{i=1}^N \|\Delta W_i - \Delta W_i(V_rV_r^\top)\|_F^2 \\ &= \frac{1}{N}\sum_{i=1}^N \|\Delta W_i(I - V_rV_r^\top)\|_F^2. \end{aligned}$$

By Lemma F.1 with $P = V_rV_r^\top$,

$$\|\Delta W_i(I - V_rV_r^\top)\|_F^2 = \|Z_i(I - V_rV_r^\top)\|_F^2.$$

Therefore

$$E_{\mathbf{PFL}}(A^\star, \{B_i^\star\}_{i=1}^N)^2 = \frac{1}{N}\sum_{i=1}^N \|Z_i(I - V_rV_r^\top)\|_F^2.$$

By Lemma F.2 with $M = I - V_rV_r^\top$,

$$\sum_{i=1}^N \|Z_i(I - V_rV_r^\top)\|_F^2 = \|\tilde{Z}(I - V_rV_r^\top)\|_F^2 = \|\tilde{Z} - \tilde{Z}V_rV_r^\top\|_F^2.$$

Hence

$$E_{\mathbf{PFL}}(A^\star, \{B_i^\star\}_{i=1}^N)^2 = \frac{1}{N}\|\tilde{Z} - \tilde{Z}V_rV_r^\top\|_F^2.$$

By the optimality of the SVD solution,

$$\|\tilde{Z} - \tilde{Z}V_r V_r^\top\|_F^2 = \min_{V_r^\top V_r = I_r} \|\tilde{Z} - \tilde{Z}V_r V_r^\top\|_F^2 = \sum_{j>r} \sigma_j^2(\tilde{Z}).$$

Defining $\mathcal{R} := \sum_{j>r} \sigma_j^2(\tilde{Z})$ gives

$$E_{\mathbf{PFL}}(A^\star, \{B_i^\star\}_{i=1}^N)^2 = \frac{1}{N}\|\tilde{Z} - \tilde{Z}V_r V_r^\top\|_F^2 = \frac{\mathcal{R}}{N}.$$

Step 2: no shared right factor does better. Fix an arbitrary $A' \in \mathbb{R}^{r\times k}$ and arbitrary $\{B_i'\}$, and let P be the orthogonal projector onto the row space of A', so that $\mathrm{rank}(P) \le r$. By Lemma F.3 applied to each client with $W = \Delta W_i$,

$$\|\Delta W_i - B_i' A'\|_F^2 \;\ge\; \min_C \|\Delta W_i - CA'\|_F^2 \;=\; \|\Delta W_i(I-P)\|_F^2.$$

By Lemma F.1 and Lemma F.2,

$$\frac{1}{N}\sum_{i=1}^N \|\Delta W_i(I-P)\|_F^2 = \frac{1}{N}\sum_{i=1}^N \|Z_i(I-P)\|_F^2 = \frac{1}{N}\|\tilde{Z} - \tilde{Z}P\|_F^2.$$

The matrix $\tilde{Z}P$ has rank at most r, so by the Eckart–Young–Mirsky theorem (Eckart & Young, 1936; Mirsky, 1960),

$$\|\tilde{Z} - \tilde{Z}P\|_F^2 \;\ge\; \min_{\mathrm{rank}(M)\le r} \|\tilde{Z} - M\|_F^2 \;=\; \sum_{j>r} \sigma_j^2(\tilde{Z}) \;=\; \mathcal{R}.$$

Combining the three displays gives $E_{\mathbf{PFL}}(A', \{B_i'\})^2 \ge \mathcal{R}/N$, and Step 1 shows that PERSIA attains this value.

Step 3: invariance. By Lemma F.1 and the block structure of $\tilde{Z}$,

$$\tilde{Z}^\top\tilde{Z} = \sum_{i=1}^N Z_i^\top Z_i = \sum_{i=1}^N \Delta W_i^\top \Delta W_i.$$

The right singular vectors of $\tilde{Z}$ are eigenvectors of $\tilde{Z}^\top\tilde{Z}$ and its squared singular values are the corresponding eigenvalues, so $\mathcal{R}$ and the top-r right singular subspace (unique whenever $\sigma_r(\tilde{Z}) > \sigma_{r+1}(\tilde{Z})$) are functions of $\{\Delta W_i\}$ alone. A reparameterization $(B_i, A_i) \mapsto (B_i T_i, T_i^{-1} A_i)$ with $T_i \in GL(r)$ leaves every ΔW_i unchanged and therefore leaves $\mathcal{R}$, the selected subspace, and every rebased update $B_i^\star A^\star = \Delta W_i V_r V_r^\top$ unchanged. □

Remark F.6 (Weighted clients)**.** *For client weights $p_i \ge 0$ with $\sum_i p_i = 1$, replacing the blocks Z_i of $\tilde{Z}$ by $\sqrt{Np_i}\,Z_i$ and repeating the proof verbatim shows that PERSIA with the weighted stack minimizes $\sum_i p_i\|\Delta W_i - B_i'A'\|_F^2$ over all shared A' of rank at most r. Choosing p_i inversely proportional to $\|Z_i\|_F^2$ equalizes the influence of clients whose updates differ in scale, for instance because their local-step budgets differ (Table 11).*

F.5 Proof of Proposition E.6

Proof. Let $S_I := \widetilde{Z}^\top\widetilde{Z}$, $S_K := \widetilde{Z}^\top K_O \widetilde{Z}$, and $E := S_K - S_I$. By Proposition E.4, $\|\overline{W}(I-P)\|_F^2 = \frac{1}{N^2}\mathrm{tr}\big(S_K(I-P)\big)$ for every orthogonal projector P. Let P_Z and P_W be the top-r eigenprojectors of S_I and S_K; P_W is also the top-r right singular projector of $\overline{W}$, since $\overline{W}^\top\overline{W} = S_K/N^2$. The lower bound $E_{\mathrm{PER}} \ge E_{\mathrm{ora}}$ holds because P_W minimizes $\mathrm{tr}(S_K(I-P))$ over rank-r projectors. For the upper bound,

$$N^2\big(E_{\mathrm{PER}} - E_{\mathrm{ora}}\big) = \mathrm{tr}\big(S_K(P_W - P_Z)\big) = \underbrace{\mathrm{tr}\big(S_I(P_W - P_Z)\big)}_{\le 0} + \mathrm{tr}\big(E(P_W - P_Z)\big)$$

$$\le \|E\|_2\,\|P_W - P_Z\|_* \;\le\; \|E\|_2\big(\|P_W\|_* + \|P_Z\|_*\big) = 2r\|E\|_2,$$

where the first term is nonpositive because P_Z maximizes $\mathrm{tr}(S_I P)$ over rank-r projectors, and we used $|\mathrm{tr}(XY)| \le \|X\|_2\|Y\|_*$. Dividing by N^2 proves the claim. If $K_O = I$, then $E = 0$. □

F.6 Proofs of the global-FL statements in Appendix E.2

Lemma F.7 (Polar factor). *Every $B_i \in \mathbb{R}^{d\times r}$ with $r \le d$ admits $O_i \in \mathbb{R}^{d\times r}$ with $O_i^\top O_i = I_r$ and $B_i = O_i G_i$, where $G_i = (B_i^\top B_i)^{1/2}$; if B_i has full column rank, $O_i = B_i G_i^{-1}$.*

Proof. Let $B_i = U\Sigma V^\top$ be a thin SVD with $U \in \mathbb{R}^{d\times r}$, $U^\top U = I_r$, and $V \in \mathbb{R}^{r\times r}$ orthogonal. Then $G_i = V\Sigma V^\top$, and $O_i := UV^\top$ satisfies $O_i^\top O_i = VU^\top UV^\top = I_r$ and $O_i G_i = UV^\top V\Sigma V^\top = U\Sigma V^\top = B_i$. If B_i has full column rank, G_i is invertible and $O_i = B_i G_i^{-1}$. □

Proof of Proposition E.4. By Lemma F.7, $\Delta W_i = O_i Z_i$, so $\overline{W} = \frac{1}{N}\sum_i O_i Z_i = \frac{1}{N}[O_1|\cdots|O_N]\widetilde{Z} = \overline{O}\widetilde{Z}$. For any P, with $Y := \widetilde{Z}(I-P)$, $\|\overline{W}(I-P)\|_F^2 = \|\overline{O}Y\|_F^2 = \mathrm{tr}(Y^\top \overline{O}^\top \overline{O} Y) = \frac{1}{N^2}\mathrm{tr}(Y^\top K_O Y)$. □

Proof of Proposition E.3. K_O is symmetric positive semidefinite with identity diagonal blocks ($O_i^\top O_i = I_r$) and off-diagonal blocks of spectral norm at most μ. For a unit vector $x = (x_1, \ldots, x_N)$ partitioned conformally, $x^\top K_O x = \sum_i \|x_i\|^2 + \sum_{i\neq j} x_i^\top O_i^\top O_j x_j \le 1 + \mu \sum_{i\neq j} \|x_i\|\|x_j\| \le 1 + (N-1)\mu$, using $\sum_{i\neq j}\|x_i\|\|x_j\| = (\sum_i \|x_i\|)^2 - \sum_i \|x_i\|^2 \le (N-1)\sum_i \|x_i\|^2$. Hence $\lambda_{\max}(K_O) \le 1 + (N-1)\mu$. By Proposition E.4 and equation 10, with $Y = \widetilde{Z}(I-P_Z)$, $E_{\mathrm{FL}}(A^\star, B^\star)^2 = \|\overline{W}(I-P_Z)\|_F^2 = \frac{1}{N^2}\mathrm{tr}(Y^\top K_O Y) \le \frac{\lambda_{\max}(K_O)}{N^2}\|Y\|_F^2 = \frac{\lambda_{\max}(K_O)}{N^2}\mathcal{R}_Z$. □

Proof of Corollary E.7. (i) is Proposition 3.1. (ii) If $O_i^\top O_j = 0$ for $i \neq j$, then $K_O = I$ and Proposition E.4 gives $\|\overline{W}(I-P)\|_F^2 = \frac{1}{N^2}\|\widetilde{Z}(I-P)\|_F^2$ for all P, so both objectives have the same minimizer. (iii) is Corollary 3.2. (iv) If $\rho = \|K_O - I\|_2 < 1$, then $(1-\rho)I \preceq K_O \preceq (1+\rho)I$, and Proposition E.4 gives the two-sided bound. With P_W the oracle projector, $E_{\mathrm{FL}}(A^\star, B^\star)^2 = \|\overline{W}(I-P_Z)\|_F^2 \le \frac{1+\rho}{N^2}\|\widetilde{Z}(I-P_Z)\|_F^2 \le \frac{1+\rho}{N^2}\|\widetilde{Z}(I-P_W)\|_F^2 \le \frac{1+\rho}{1-\rho}\|\overline{W}(I-P_W)\|_F^2$. (v) Let $S := \overline{W}^\top \overline{W}$, so that P_W is its top-r eigenprojector and $S = P_W S P_W + (I-P_W)S(I-P_W)$. Then $E_{\mathrm{PER}} - E_{\mathrm{ora}} = \mathrm{tr}(SP_W) - \mathrm{tr}(SP_Z) \le \mathrm{tr}(P_W S P_W) - \mathrm{tr}(P_W S P_W P_Z) = \mathrm{tr}\big(P_W S P_W (I - P_Z)\big) \le \lambda_1(S)\,\|(I-P_Z)P_W\|_F^2$, and $\|(I-P_Z)P_W\|_F = \|\sin\Theta(V_r^W, V_r^Z)\|_F$. For the second bound, $N^2 S = \widetilde{Z}^\top K_O \widetilde{Z} = \widetilde{Z}^\top \widetilde{Z} + E$ with $E := \widetilde{Z}^\top(K_O - I)\widetilde{Z}$, and the argument of Proposition E.3 applied to $K_O - I$ (zero diagonal blocks) gives $\|E\|_2 \le (N-1)\mu\,\sigma_1^2(\widetilde{Z})$. If $2\|E\|_2 < \delta_Z$, Weyl's inequality separates the top-r eigenvalues of $N^2 S$ from the rest by more than $\delta_Z/2$, and the Davis–Kahan $\sin\Theta$ theorem (Davis & Kahan, 1970) gives $\|\sin\Theta\|_F \le \sqrt{r}\,\|E\|_2/(\delta_Z/2)$. Squaring and multiplying by $\lambda_1(S) = \sigma_1^2(\overline{W})$ gives the stated bound. □

F.7 Proof of Corollary: Eliminating $GL(\cdot)$-symmetry error

Corollary F.8 (Eliminating $GL(\cdot)$-symmetry error)**.** *Suppose all clients learn the same functionality under various symmetrical representations.* $\Delta W_i = W^* = BA = (BT_i)(T_i^{-1}A)$*, where* $T_i \in \mathbb{R}^{r\times r}$ *is a client-specific invertible transformation, or, more generally, that* $\dim\big(\sum_i \mathrm{row}(\Delta W_i)\big) \le r$*. We can show that the PERSIA factorization is equivalent to the ideal aggregation in both settings:*

$$E_{\mathbf{PFL}}(A^*, \{B_i^*\}_{i=1}^N) := \Big(\frac{1}{N}\sum_{i=1}^N \|\Delta W_i - B_i^* A^*\|_F^2\Big)^{1/2} = 0,$$
$$E_{\mathbf{FL}}(A^*, B^*) := \|\overline{W} - B^* A^*\|_F = 0.$$

Proof. **Step 1: the stack has rank at most r.** By Lemma F.1, $Z_i^\top Z_i = \Delta W_i^\top \Delta W_i$, so Z_i and ΔW_i have the same null space and hence the same row space. The row space of $\tilde{Z}$ is the sum of the row spaces of its blocks,

$$\mathrm{row}(\tilde{Z}) = \sum_{i=1}^N \mathrm{row}(Z_i) = \sum_{i=1}^N \mathrm{row}(\Delta W_i),$$

whose dimension is at most r by assumption. Under the symmetry assumption, $\mathrm{row}(\Delta W_i) = \mathrm{row}(BA)$ for every i, which has dimension at most r. Hence $\mathrm{rank}(\tilde{Z}) \le r$ and $\mathcal{R} = \sum_{j>r} \sigma_j^2(\tilde{Z}) = 0$.

Step 2: personalized error. By Proposition 3.1, $E_{\mathbf{PFL}}(A^*, \{B_i^*\})^2 = \mathcal{R}/N = 0$. Equivalently, the columns of V_r span a subspace containing $\mathrm{row}(\tilde{Z})$, so $\Delta W_i V_r V_r^\top = \Delta W_i$ and $B_i^* A^* = \Delta W_i$ for every i.

Step 3: global error. By Proposition E.2, $E_{\mathbf{FL}}(A^*, B^*)^2 \le \mathcal{R}/N = 0$. Equivalently, by equation 10 and Step 2,

$$B^* A^* = \overline{W} V_r V_r^\top = \frac{1}{N}\sum_{i=1}^N \Delta W_i V_r V_r^\top = \frac{1}{N}\sum_{i=1}^N \Delta W_i = \overline{W}.$$

No rank condition on B or on the B_i is required. □

G PROOF OF THEOREMS

This appendix proves Theorem 4.4. The proof separates the local SGD descent from the change of the objective caused by aggregation, bounds the latter by the aggregation displacement, and bounds the displacement by the PERSIA residual.

G.1 SETTING AND NOTATION

We consider one adapted layer; with several layers, x collects all factor pairs, $\|\cdot\|$ is the Euclidean norm over all of them, and the residuals are summed over layers. Write $x = (A, B)$ with $A \in \mathbb{R}^{r\times k}$, $B \in \mathbb{R}^{d\times r}$, and $\|x\|^2 := \|A\|_F^2 + \|B\|_F^2$. For client i,

$$F_i(A, B) := f_i(W_0 + BA),$$

$$\nabla_A F_i(A, B) = B^\top \nabla f_i(W_0 + BA), \qquad \nabla_B F_i(A, B) = \nabla f_i(W_0 + BA)\, A^\top,$$

which follows from $\mathrm{d}(BA) = (\mathrm{d}B)A + B(\mathrm{d}A)$ and $\langle M, (\mathrm{d}B)A\rangle = \langle MA^\top, \mathrm{d}B\rangle$, $\langle M, B\,\mathrm{d}A\rangle = \langle B^\top M, \mathrm{d}A\rangle$. The factor-stationarity measure is $\mathcal{G}_i(x) := \|\nabla_A F_i(x)\|_F^2 + \|\nabla_B F_i(x)\|_F^2 = \|\nabla F_i(x)\|^2$.

Protocol. In round t, client i starts from $x_i^{t,0} = (A^t, B_i^t)$ in personalized FL and from $x_i^{t,0} = x^t = (A^t, B^t)$ in global FL. For $s = 0, \dots, \tau - 1$ it draws a stochastic gradient $g_i^{t,s} := g_i(W_0 + B_i^{t,s}A_i^{t,s}; \xi_i^{t,s})$ and sets

$$x_i^{t,s+1} = x_i^{t,s} - \eta\, \hat{g}_i^{t,s}, \qquad \hat{g}_i^{t,s} := \big((B_i^{t,s})^\top g_i^{t,s},\ g_i^{t,s}(A_i^{t,s})^\top\big).$$

Aggregation maps $\{x_i^{t,\tau}\}_i$ to $\{x_i^{t+1,0}\}_i$. Since the broadcast $A^\star = V_r^\top$ has orthonormal rows, Assumption 4.3 requires $\rho_A \geq 1$. We write $\Delta W_i^{t,s} := B_i^{t,s}A_i^{t,s}$, $\mathcal{F}^{t,s}$ for the history up to the draw of $\xi^{t,s}$, and $\mathbb{E}_{t,s}$ for the conditional expectation given $\mathcal{F}^{t,s}$. The samples $\xi_i^{t,s}$ are drawn independently of $\mathcal{F}^{t,s}$, so Assumption 4.2 holds conditionally on $\mathcal{F}^{t,s}$. The initialization is fixed (or, equivalently, all statements are conditional on it), so D is deterministic.

G.2 CONSEQUENCES OF THE ASSUMPTIONS

Lemma G.1 (Consequences of Assumptions 4.1 to 4.3)**.** *Let $\mathcal{X} := \{(A, B) : \|A\|_2 \leq \rho_A,\ \|B\|_2 \leq \rho_B\}$ and define $G_\nabla, \Gamma^2, \sigma_F^2, L_F$ as in Section 4. For all $x = (A, B) \in \mathcal{X}$ and every client i:*

(a) $\|BA\|_F \leq \rho_A\rho_B\sqrt{r}$ and $\|\nabla f_i(W_0 + BA)\|_F \leq G_\nabla$;

(b) $\mathbb{E}\|g_i(W_0 + BA; \xi)\|_F^2 \leq \Gamma^2$;

(c) the factor estimate $\hat{g} = (B^\top g_i,\ g_i A^\top)$ satisfies $\mathbb{E}[\hat{g}] = \nabla F_i(x)$ and $\mathbb{E}\|\hat{g} - \nabla F_i(x)\|^2 \leq \sigma_F^2$;

(d) $\|\nabla F_i(x) - \nabla F_i(x')\| \leq L_F\|x - x'\|$ for all $x, x' \in \mathcal{X}$;

(e) F_i attains its minimum on $\mathcal{X}$, so $D < \infty$.

Proof. **(a)** BA has rank at most r, so $\|BA\|_F \leq \sqrt{r}\|BA\|_2 \leq \sqrt{r}\|B\|_2\|A\|_2 \leq \rho_A\rho_B\sqrt{r}$. By Assumption 4.1, $\|\nabla f_i(W_0 + BA)\|_F \leq \|\nabla f_i(W_0)\|_F + L\|BA\|_F \leq G_\nabla$.

(b) By Assumption 4.2 the cross term vanishes in expectation, so $\mathbb{E}\|g_i\|_F^2 = \|\nabla f_i\|_F^2 + \mathbb{E}\|g_i - \nabla f_i\|_F^2 \leq G_\nabla^2 + \sigma^2 = \Gamma^2$.

(c) Unbiasedness follows from linearity of $g \mapsto (B^\top g, gA^\top)$. For the variance, using $\|B^\top M\|_F \leq \|B\|_2\|M\|_F$ and $\|MA^\top\|_F \leq \|A\|_2\|M\|_F$,

$$\mathbb{E}\|\hat{g} - \nabla F_i(x)\|^2 = \mathbb{E}\|B^\top(g_i - \nabla f_i)\|_F^2 + \mathbb{E}\|(g_i - \nabla f_i)A^\top\|_F^2 \leq (\rho_B^2 + \rho_A^2)\sigma^2 = \sigma_F^2.$$

(d) Let $x = (A, B)$, $x' = (A', B') \in \mathcal{X}$, $M := \nabla f_i(W_0 + BA)$, $M' := \nabla f_i(W_0 + B'A')$, $a := \|A - A'\|_F$, and $b := \|B - B'\|_F$. Then

$$\|BA - B'A'\|_F = \|(B - B')A + B'(A - A')\|_F \leq \rho_A b + \rho_B a,$$

$$\|M - M'\|_F \leq L(\rho_A b + \rho_B a), \qquad \|M\|_2 \leq \|M\|_F \leq G_\nabla.$$

Therefore

$$\begin{aligned}\|\nabla_A F_i(x) - \nabla_A F_i(x')\|_F &= \|(B-B')^\top M + B'^\top (M-M')\|_F \\ &\le G_\nabla b + L\rho_B(\rho_A b + \rho_B a) = c\,b + L\rho_B^2 a, \\ \|\nabla_B F_i(x) - \nabla_B F_i(x')\|_F &= \|M(A-A')^\top + (M-M')A'^\top\|_F \\ &\le G_\nabla a + L\rho_A(\rho_A b + \rho_B a) = c\,a + L\rho_A^2 b,\end{aligned}$$

with $c := G_\nabla + L\rho_A\rho_B$. By the triangle inequality in $\mathbb{R}^2$,

$$\begin{aligned}\|\nabla F_i(x) - \nabla F_i(x')\| &\le \big\|(c\,b,\, c\,a)\big\|_2 + \big\|(L\rho_B^2 a,\, L\rho_A^2 b)\big\|_2 \\ &\le \big(c + L\max\{\rho_A^2, \rho_B^2\}\big)\|x - x'\| \le L_F\|x - x'\|,\end{aligned}$$

since $\rho_A\rho_B + \max\{\rho_A^2, \rho_B^2\} \le (\rho_A + \rho_B)^2$.

(e) F_i is continuous and $\mathcal{X}$ is compact. □

Since $\mathcal{X}$ is convex and every iterate lies in $\mathcal{X}$ by Assumption 4.3, the descent inequality $F_i(y) \le F_i(x) + \langle \nabla F_i(x), y - x\rangle + \frac{L_F}{2}\|y - x\|^2$ holds for consecutive iterates x, y.

G.3 Local descent and the master inequality

Lemma G.2 (Local SGD descent). *If $0 < \eta \le 1/L_F$, then for every client, round, and local step,*

$$\mathbb{E}_{t,s} F_i(x_i^{t,s+1}) \le F_i(x_i^{t,s}) - \frac{\eta}{2}\mathcal{G}_i(x_i^{t,s}) + \frac{L_F\eta^2}{2}\sigma_F^2.$$

Proof. Apply the descent inequality with $x = x_i^{t,s}$ and $y = x_i^{t,s+1} = x - \eta\hat{g}_i^{t,s}$:

$$F_i(x_i^{t,s+1}) \le F_i(x_i^{t,s}) - \eta\langle \nabla F_i(x_i^{t,s}), \hat{g}_i^{t,s}\rangle + \frac{L_F\eta^2}{2}\|\hat{g}_i^{t,s}\|^2.$$

By Lemma G.1(c), $\mathbb{E}_{t,s}\hat{g}_i^{t,s} = \nabla F_i(x_i^{t,s})$ and $\mathbb{E}_{t,s}\|\hat{g}_i^{t,s}\|^2 \le \mathcal{G}_i(x_i^{t,s}) + \sigma_F^2$. Taking $\mathbb{E}_{t,s}$,

$$\mathbb{E}_{t,s} F_i(x_i^{t,s+1}) \le F_i(x_i^{t,s}) - \eta\Big(1 - \frac{L_F\eta}{2}\Big)\mathcal{G}_i(x_i^{t,s}) + \frac{L_F\eta^2}{2}\sigma_F^2,$$

and $1 - L_F\eta/2 \ge 1/2$ for $\eta \le 1/L_F$. □

Definition G.3 (Aggregation jump). *The average change of the client objectives caused only by the round-t aggregation is*

$$J_t := \frac{1}{N}\sum_{i=1}^N \Big(F_i(x_i^{t+1,0}) - F_i(x_i^{t,\tau})\Big), \qquad (J_t)_+ := \max\{J_t, 0\}.$$

Lemma G.4 (Master inequality for any aggregation rule). *If $0 < \eta \le 1/L_F$ and every iterate lies in $\mathcal{X}$, then for any aggregation rule,*

$$\frac{1}{NT\tau}\sum_{i=1}^N\sum_{t=0}^{T-1}\sum_{s=0}^{\tau-1}\mathbb{E}\,\mathcal{G}_i(x_i^{t,s}) \le \frac{2D}{\eta\tau T} + L_F\eta\sigma_F^2 + \frac{2}{\eta\tau T}\sum_{t=0}^{T-1}\mathbb{E}(J_t)_+.$$

Proof. Take full expectations in Lemma G.2, sum over $s = 0, \dots, \tau - 1$, and average over clients:

$$\frac{1}{N}\sum_i \mathbb{E}F_i(x_i^{t,\tau}) \le \frac{1}{N}\sum_i \mathbb{E}F_i(x_i^{t,0}) - \frac{\eta}{2N}\sum_{i,s}\mathbb{E}\,\mathcal{G}_i(x_i^{t,s}) + \frac{\tau L_F\eta^2}{2}\sigma_F^2.$$

By Definition G.3, $\frac{1}{N}\sum_i \mathbb{E}F_i(x_i^{t+1,0}) = \frac{1}{N}\sum_i \mathbb{E}F_i(x_i^{t,\tau}) + \mathbb{E}J_t \le \frac{1}{N}\sum_i \mathbb{E}F_i(x_i^{t,\tau}) + \mathbb{E}(J_t)_+$. Chaining the two inequalities and summing over $t = 0, \dots, T-1$ telescopes the objective values:

$$\frac{\eta}{2N}\sum_{i,t,s}\mathbb{E}\,\mathcal{G}_i(x_i^{t,s}) \le \frac{1}{N}\sum_i\Big(F_i(x_i^{0,0}) - \mathbb{E}F_i(x_i^{T,0})\Big) + \frac{T\tau L_F\eta^2}{2}\sigma_F^2 + \sum_{t=0}^{T-1}\mathbb{E}(J_t)_+.$$

Since $x_i^{T,0} \in \mathcal{X}$, $F_i(x_i^{T,0}) \ge \min_{\mathcal{X}} F_i$, so the first term is at most D. Dividing by $\eta\tau T/2$ gives the claim. □

G.4 BOUNDING THE AGGREGATION JUMP

Lemma G.5 (Jump bound). *Let $\delta_i^t := \Delta W_i^{t+1,0} - \Delta W_i^{t,\tau}$ be the change of client i's induced update caused by aggregation, and $Q_t := \mathbb{E}\big[\frac{1}{N}\sum_i \|\delta_i^t\|_F^2\big]$. Then*

$$\mathbb{E}(J_t)_+ \le G_\nabla \sqrt{Q_t} + \frac{L}{2} Q_t.$$

Proof. Since $F_i(x) = f_i(W_0 + \Delta W)$, Assumption 4.1 gives

$$F_i(x_i^{t+1,0}) - F_i(x_i^{t,\tau}) \le \langle \nabla f_i(W_0 + \Delta W_i^{t,\tau}), \delta_i^t \rangle + \frac{L}{2}\|\delta_i^t\|_F^2 \le G_\nabla \|\delta_i^t\|_F + \frac{L}{2}\|\delta_i^t\|_F^2,$$

using Cauchy–Schwarz and Lemma G.1(a). The right-hand side is nonnegative, so averaging over clients bounds $(J_t)_+$ as well. Taking expectations and applying Cauchy–Schwarz over clients and Jensen's inequality to the concave square root,

$$\mathbb{E}\Big[\frac{1}{N}\sum_i \|\delta_i^t\|_F\Big] \le \mathbb{E}\Big[\Big(\frac{1}{N}\sum_i \|\delta_i^t\|_F^2\Big)^{1/2}\Big] \le \sqrt{Q_t},$$

which proves the claim. □

G.5 THE PERSIA RESIDUAL

Definition G.6 (Per-round PERSIA residual). *At the end of round t, let $G_i^t := \big((B_i^{t,\tau})^\top B_i^{t,\tau}\big)^{1/2}$, $Z_i^t := G_i^t A_i^{t,\tau}$, let $\tilde{Z}^t$ stack the Z_i^t, and let V_r^t hold its top-r right singular vectors, $P_r^t := V_r^t (V_r^t)^\top$. Define*

$$\mathcal{R}^t := \|\tilde{Z}^t - \tilde{Z}^t P_r^t\|_F^2 = \sum_{j>r} \sigma_j^2(\tilde{Z}^t), \qquad \bar{\mathcal{R}}_t := \frac{\mathbb{E}[\mathcal{R}^t]}{N}.$$

Lemma G.7 (PFL displacement equals the residual). *In personalized FL, $\delta_i^t = -\Delta W_i^{t,\tau}(I - P_r^t)$ and $Q_t = \bar{\mathcal{R}}_t$.*

Proof. By equation 7, $\Delta W_i^{t+1,0} = B_i^\star A^\star = \Delta W_i^{t,\tau} P_r^t$, so $\delta_i^t = -\Delta W_i^{t,\tau}(I - P_r^t)$. By Step 1 of the proof of Proposition F.5, $\frac{1}{N}\sum_i \|\Delta W_i^{t,\tau}(I - P_r^t)\|_F^2 = \mathcal{R}^t/N$ pathwise; taking expectations gives $Q_t = \bar{\mathcal{R}}_t$. □

Lemma G.8 (Residual bound from a common right factor). *If all clients start round t from the same right factor, $A_i^{t,0} = A^t$ for all i, then*

$$\bar{\mathcal{R}}_t \le \rho_B^4 \Gamma^2 \eta^2 \tau^2 = \Phi\, \eta^2 \tau^2.$$

Proof. **Step 1: comparison with a rank-r stack.** By the Eckart–Young–Mirsky theorem, $\mathcal{R}^t = \min_{\mathrm{rank}(M)\le r} \|\tilde{Z}^t - M\|_F^2$. The matrix

$$M^t := \begin{bmatrix} G_1^t A^t \\ \vdots \\ G_N^t A^t \end{bmatrix}$$

has all its rows in the row space of $A^t \in \mathbb{R}^{r\times k}$, so $\mathrm{rank}(M^t) \le r$. Hence, by Lemma F.2,

$$\mathcal{R}^t \le \|\tilde{Z}^t - M^t\|_F^2 = \sum_{i=1}^N \|G_i^t(A_i^{t,\tau} - A^t)\|_F^2 \le \sum_{i=1}^N \|G_i^t\|_2^2 \, \|A_i^{t,\tau} - A^t\|_F^2 \le \rho_B^2 \sum_{i=1}^N \|A_i^{t,\tau} - A^t\|_F^2,$$

where $\|G_i^t\|_2^2 = \|(B_i^{t,\tau})^\top B_i^{t,\tau}\|_2 = \|B_i^{t,\tau}\|_2^2 \le \rho_B^2$.

Step 2: drift of the right factor. Since $A_i^{t,0} = A^t$,

$$A_i^{t,\tau} - A^t = -\eta \sum_{s=0}^{\tau-1} (B_i^{t,s})^\top g_i^{t,s}.$$

Using $\|\sum_{s=0}^{\tau-1} u_s\|_F^2 \le \tau \sum_{s=0}^{\tau-1} \|u_s\|_F^2$ and $\|(B_i^{t,s})^\top g_i^{t,s}\|_F \le \rho_B \|g_i^{t,s}\|_F$,

$$\|A_i^{t,\tau} - A^t\|_F^2 \le \eta^2 \tau \rho_B^2 \sum_{s=0}^{\tau-1} \|g_i^{t,s}\|_F^2.$$

By the tower property and Lemma G.1(b), $\mathbb{E}\|g_i^{t,s}\|_F^2 = \mathbb{E}\big[\mathbb{E}_{t,s}\|g_i^{t,s}\|_F^2\big] \le \Gamma^2$, so $\mathbb{E}\|A_i^{t,\tau} - A^t\|_F^2 \le \rho_B^2 \Gamma^2 \eta^2 \tau^2$.

Step 3: combine. Taking expectations in Step 1 and inserting Step 2,

$$\mathbb{E}[\mathcal{R}^t] \le \rho_B^2 \sum_{i=1}^{N} \rho_B^2 \Gamma^2 \eta^2 \tau^2 = N \rho_B^4 \Gamma^2 \eta^2 \tau^2.$$

Dividing by N gives the claim. □

G.6 Proof of Theorem 4.4

Proof. **Part (i).** By Lemmas G.5 and G.7, $\mathbb{E}(J_t)_+ \le G_\nabla \sqrt{\bar{\mathcal{R}}_t} + \frac{L}{2}\bar{\mathcal{R}}_t$. Inserting this into Lemma G.4 gives equation 8. No bound on the size of $\mathcal{R}^t$ is used.

Part (ii). In personalized FL every client starts each round from the broadcast A^t, so Lemma G.8 gives $\bar{\mathcal{R}}_t \le \Phi \eta^2 \tau^2$ and

$$\frac{2}{\eta\tau T} \sum_{t=0}^{T-1} \Big(G_\nabla \sqrt{\bar{\mathcal{R}}_t} + \frac{L}{2}\bar{\mathcal{R}}_t \Big) \le \frac{2}{\eta\tau} \Big(G_\nabla \eta\tau \sqrt{\Phi} + \frac{L}{2}\eta^2\tau^2\Phi \Big) = 2G_\nabla \sqrt{\Phi} + L\eta\tau\Phi.$$

Therefore, by part (i),

$$\frac{1}{NT\tau} \sum_{i,t,s} \mathbb{E}\, \mathcal{G}_i(x_i^{t,s}) \le 2G_\nabla \sqrt{\Phi} + h(\eta), \qquad h(\eta) := \frac{2D}{\eta\tau T} + \eta H, \qquad H = L_F \sigma_F^2 + L\Phi\tau. \tag{14}$$

The unconstrained minimizer of h is $\eta_\star = \sqrt{2D/(\tau T H)}$, with $h(\eta_\star) = 2\sqrt{2DH/(\tau T)}$.

Case $\eta_\star \le 1/L_F$. Then $\eta = \eta_\star$ and $h(\eta) = 2\sqrt{2DH/(\tau T)}$.

Case $\eta_\star > 1/L_F$. Then $\eta = 1/L_F$ and $H < 2DL_F^2/(\tau T)$, i.e., $\sqrt{H} < L_F \sqrt{2D/(\tau T)}$. Hence

$$h(1/L_F) = \frac{2DL_F}{\tau T} + \frac{H}{L_F} = \frac{2DL_F}{\tau T} + \sqrt{H}\,\frac{\sqrt{H}}{L_F} < \frac{2DL_F}{\tau T} + \sqrt{\frac{2DH}{\tau T}}.$$

In both cases $h(\eta) \le 2\sqrt{2DH/(\tau T)} + 2DL_F/(\tau T)$, which together with equation 14 proves equation 9. □

Remark G.9 (Properties of the bound)**.** *(a)* Floor. *The term $2G_\nabla\sqrt{\Phi} = 2G_\nabla \rho_B^2 \Gamma$ originates from the first-order part $G_\nabla \sqrt{\bar{\mathcal{R}}_t} \le G_\nabla \eta\tau\sqrt{\Phi}$ of the jump, which is of the same order $\eta\tau$ as the local progress per round and therefore survives the division by $\eta\tau$. It does not depend on T, η, or τ. (b)* Monotonicity in τ. *Since $H/\tau = L_F\sigma_F^2/\tau + L\Phi$, the right-hand side of equation 9 equals $2G_\nabla\sqrt{\Phi} + 2\sqrt{2DL_F\sigma_F^2/(\tau T) + 2DL\Phi/T} + 2DL_F/(\tau T)$, which is non-increasing in τ for fixed T (for constants that hold uniformly in τ). (c)* Zero residual. *If $\mathcal{R}^t = 0$ in every round, as in Corollary 3.2, part (i) with $\eta = \min\{1/L_F, \sqrt{2D/(\tau T L_F \sigma_F^2)}\}$ (for $\sigma > 0$) gives $2\sqrt{2DL_F\sigma_F^2/(\tau T)} + 2DL_F/(\tau T)$, the floor-free rate of SGD on each client objective over τT steps (without a speed-up in N). (d)* Other aggregation rules. *Lemmas G.4 and G.5 hold for any rule. For any rule that outputs a shared right factor of rank at most r and arbitrary client left factors, Proposition 3.1 implies $\frac{1}{N}\sum_i \|\delta_i^t\|_F^2 \ge \mathcal{R}^t/N$ pathwise for the same pre-aggregation iterates, so PERSIA minimizes the quantity Q_t that controls the upper bound on the aggregation term in every round (the jump J_t itself is not claimed to be minimal). Part (ii), in contrast, is not specific to PERSIA: for factor averaging, $\delta_i^t = B_i^{t,\tau}(\bar{A}^{t,\tau} - A_i^{t,\tau})$ with $\bar{A}^{t,\tau} := \frac{1}{N}\sum_j A_j^{t,\tau}$, and the same steps give $Q_t \le \rho_B^2 \frac{1}{N}\sum_i \mathbb{E}\|A_i^{t,\tau} - A^t\|_F^2 \le \Phi\eta^2\tau^2$. (f)* When the floor is informative. *On $\mathcal{X}$, $\mathcal{G}_i \le \|B\|_2^2\|\nabla f_i\|_F^2 + \|A\|_2^2\|\nabla f_i\|_F^2 \le (\rho_A^2 + \rho_B^2)G_\nabla^2$ trivially. Since $\Gamma \ge G_\nabla$, the floor $2G_\nabla\rho_B^2\Gamma$*

is below this trivial level only if $\rho_A^2 > \rho_B^2(2\Gamma/G_\nabla - 1)$, i.e., when the left factors are small relative to the right factor, and it is never below $2\rho_B^2 G_\nabla^2 \geq 2\sup_{\mathcal{X}} \|\nabla_A F_i\|_F^2$, so its content concerns the component $\nabla_B F_i = \nabla f_i A^\top$. (g) Scope of Assumption 4.3. *Almost-sure spectral bounds along the whole trajectory, including the rebased $B_i^\star = \Delta W_i V_r$ (for which only $\|B_i^\star\|_2 \leq \rho_A \rho_B$ is automatic), in practice correspond to bounded gradient noise and small step sizes, or to an explicit projection onto $\mathcal{X}$. (e)* Factor versus ambient stationarity. $\mathcal{G}_i(A, B) = 0$ *does not imply $\nabla f_i(W_0 + BA) = 0$; for example, both factor gradients vanish at $(A, B) = (0, 0)$ for any $\nabla f_i(W_0)$. The theorem is therefore a statement about the factor variables that the algorithm updates.*